Abstracts of the First Asia Pacific Corpus Linguistics Conference

Edited by

Michael Barlow

Helen Basturkmen

Qing Li

Table of Contents

Chen, Howard & Yang, Christine	*Developing a Chinese as Second Language Learner Corpus and a Web-based Concordancer*	
Chen, Wallace	*Building a Translation Learner Corpus: Methodology and Application*	
Cheng, Winnie	*Exploring the Limits of Phraseology*	
Choi, Incheol	*A Learner Corpus Study of Korean EFL Learners' Overgeneralization Errors*	
Collins, Peter, Borlongan, Ariane & Yao, Xinyue	*Modality in Philippine English: a diachronic study*	
Collins, Peter & Yao, Xinyue	*Recent change in non-present perfect constructions in British and American English*	
Dash, Niladri Sekhar; Dutta Chowdhury, Payel & De, Amrita	*Rules for POS Tagging of the Bengali Text Corpus*	
Dash, Niladri Sekhar & Topdar, Baisakhi	*Lexical Generativity of Bengali Prefixes: A Corpus-based Investigation*	
Demeter, Gusztav	*Co-constructed and Negotiated Apologies. Contributions of Corpus Linguistics to the Study of Speech Acts*	
Doi, Kosuke	*What Linguistic Features Determine the CEFR reading levels?*	
Doyle, Paul	Visualizing a Corpus of Spoken Classroom Discourse	
Eckart, Thomas，Quasthoff，Uwe & Goldhahn，Dirk	*Language Statistics-Based Quality Assurance for Large Corpora*	
Facchinetti, Roberta	*Ideology in News Reporting: A Multimodal Corpus-based Study*	
Fu, Xitao	*"110" in China: A Cognitive Approach*	
Gaskell, Delian	*Evaluation in Spoken Discourse: An Exploratory Investigation of Stance in the MICASE*	
Grant, Lynn& Crezee, Ineke	*Are interpreters missing the plot? Developing a corpus of 'realia' television programmes to test interpreters' knowledge of idiomatic language*	
Keisuke Hashimoto	*The revision of the collocations in some entries of タイ日大辞典 with ThaiWaC*	
Haugh, Michael; Cassidy, Steve & Peters, Pam	*Towards a Multimodal Australian National Corpus*	
He, Anping	*A study of stance in multi-type phraseology*	
Huang, Dora Zeping	*Can Concordance Improve the Lexico-grammatical Use of Abstract Nouns in L2 Writing?*	
Taku Kaneta	*Criterial features of L2 acquisition stages: contrastive interlanguage analysis at elementary level*	
Kang, Heng-ming Carlos	Gerund and Infinitive after *Like*: An Investigation of *Like to Say* and *Like Saying*	
Kitao, Kenji & Kitao, S. Kathleen	*Developing and Analyzing Corpora from Subtitles*	
Kitao, S. Kathleen & Kitao, Kenji	*A Corpus-Based Study of University Entrance Exams*	

Kovář,Vojtěch& McCarthy, Diana	*New Learner Corpus Functionality in the Sketch Engine*	
Kwary, Deny A.& Kilgarriff, Adam	*Using a Web Corpus in Indonesian Language Learning*	
Kwong, Oi Yee	*Analysing Story Structure with a Bilingual Corpus of Aesop's Fables*	
Lai, Shu-Li	*EFL Students' Perceptions of Corpus-Tools as Writing Aids*	
Lam, Thi Hoa Binh	*Corpus Linguistics – A Trend in Compiling ESP Documents for College and University Students in Vietnam*	
Le, Van	*Building and Using English-Vietnamese Parallel Corpora (EVPC)*	
Lee, Dong Ju	*Corpora and Data-Driven Learning (DDL) Approach*	
Li, Jen-I, Hsieh, Miao-Ling & Chen, Hao-Jan	L2 Acquisition of Adverb *Cai* in Mandarin Chinese	
Lim, Jooyoung & Kim, Jong-Bok	*English Measure Noun Phrases: A Usage-based Approach*	
Lin, Weiyan	*On Chinese University Students' English Metadiscoursal Oral Chunks*	
Lin, Ya-Hui, Chen, Feng-Yi，Hsieh, Miao-Ling & Fang, Shu-hua	*The Construction of Interlanguage Corpus of CSL Learners*	
Lingle, Will	*Abu Ghraib, Guantánamo, and the New York Times: A Corpus-assisted Critical Discourse Analysis*	
Masuda, Masanobu	*Linguistic Devices to Elicit Elaborated Talk in Interview Dialogues: The Case of the Corpus of Spontaneous Japanese*	
Murcia Bielsa, Susana & O'Donnell, Mick	*Using a Learner Corpus to Develop an EFL Grammar Teaching Curriculum*	
Nam, Daehyeon	*A Corpus-Based Analysis of College ESL Students' Argumentative Writing*	
Nation, Paul	*How much input do learners need? A corpus-based answer*	
Nesi, Hilary	*A New Wordtree Corpus Interface*	
Ng, Serene S.H.	*An Account of the English Loanwords in Singapore Chinese – A Corpus-Based Approach*	
Nomura, Mariko	*Interlanguage Variation in Past Tense Marking in Japanese EFL Learners' Spoken and Written Corpora*	
Okada, Akira	*Historical Analyses of English Negative Affixes from Middle English to Present-day English*	
Rocha, Marco	*Anaphoric demonstratives: a cross-linguistic corpus-based approach*	
Someya,Yasumasa, Kikuchi, Atsuko, Akasegawa, Shiro & Yamaoka, Yoichi	*Construction of a Large-scale Translation Corpus and its Re search and Pedagogical*	

	Implications	
Sung, Li-May	*Grammar of HAN 'say; said' in Sakizaya: a corpus-based perspective*	
Tanangkingsing, Michael	*Pronouns in Discourse and Word Order Change in Cebuano*	
Tasanameelarp, Asama & Laohawiriyanon, Chonlada	*Do Corpora Work for All Levels? DDL with High and Low Proficiency learners*	
Tegge, Friederike A. G.	*The Lexical Profile of Songs Used in the English Language Classroom*	
Vine, Elaine W.	*Patterns of Use of Category Ambiguous Words in Corpora and Coursebooks*	
Wei, Naixing & Zhou, Junying	*A Contrastive Study of Corresponding Lexical Items in English and Chinese*	
Xu, Qi; YIN, Ming & Mak, Winfred Wing Fung	*What Can MAKE Tell us About Second Language Learning? – A Corpus-based Study of the Verb Complementation Patterns for MAKE*	
Yang, Suxiang	*Organizational Framework in English: The Case of "It is * that"*	

Applying Corpus Linguistics across Disciplines:

A Case Study from Civil Engineering

Susan M. Conrad

Department of Applied Linguistic, Portland State University
Portland, Oregon USA

Keywords**:** writing, engineering, error analysis, multi-dimensional analysis

Introduction

The purpose of this paper is two-fold: (1) to describe findings from a corpus-based study that compares the writing of civil engineering students and practitioners, and uses the results to design materials to improve students' writing skills, and (2) to discuss this research as a case study for some of the challenges that corpus linguists face as they work with fields very different from language-related disciplines.

The corpus study was designed to address a continuing problem in engineering education:the mismatch between the writing skills of engineering program graduates and the demands of writing in the workplace\. For years this problem has been emphasized in surveys of employers and alumni of even strong engineering programs (e.g. Reave, 2004). However, the problem has typically been investigated with case studies of individual writers or courses (e.g. Winsor, 1996; Blakeslee, 2001) or with surveys of the types of communication engineers report using (Tenopir and King, 2004). No study had sought to compare the features of writing produced by a large number of practitioners and students. A corpus-based approach thus seemed especially useful for investigating the problem and identifying areas for additional writing instruction. Challenged by having no personal background in engineering, but benefitting from the collaboration of engineering practitioners and faculty, I conducted a project to collect and analyze writing from undergraduate students in numerous civil engineering classes at Portland State University and from numerous civil engineering firms in the local area.

The Study

The corpus for the project currently consists of approximately 350 documents from 10 firms in the local community and 400 papers from undergraduate students in 19 classes, covering a range of sub-fields and document types. Writers include both first and second language speakers of English. Analyses have been conducted for numerous concerns in writing, including rhetorical organization. For this paper, I focus on three corpus-based analyses of sentence-level concerns. These analyses were conducted on reports and technical memoranda, two genres written by both students and practitioners:

- an error analysis, focused on non-standard use of English grammar and punctuation

- one dimension from a multi-dimensional analysis, based on Biber's (1988) dimension of "Impersonal Style"
- an analysis of sentence structure, comparing the use of complex or embedded structures vs simple sentences.

The three analyses demonstrate different ways of using the corpus. The first required inserting error codes by hand into the corpus, and then counting the codes with an automatic program. The second was conducted with specially written programs using a tagged version of the corpus. The third was a follow-up to the multi-dimensional analysis and was conducted entirely with hand counts of a sample of sentences from the corpus.

The findings from the analyses show that students produce statistically significantly more errors than the practitioners, and that even though errors decrease for senior-level papers written (and edited) by groups of students, the errors continue to be significantly more common than in practitioner writing. The multi-dimensional analysis and the sentence structure analysis both find statistically significantly more subordinate structures in the student writing, often to the point of obscuring meaning. Frequently, these problems with sentence structure are compounded by a high frequency of passive voice constructions.

In addition to analyses of the corpus, the study also includes interviews with students, practitioners, and faculty. Most notably, the interviews have found that practitioners are keenly aware of a connection between engineering and communication: specifically, that writing requires just as much accuracy and precision as engineering calculations do. Complex sentences, ineffective passives, and even non-standard grammar are all seen to contribute to imprecision. However, faculty and students expressed no such awareness of the integration of engineering and writing, seeing writing as a skill adjunct to engineering practice. Nevertheless, students and faculty both appear eager to use the writing materials being developed out of the project, examples of which are shared in the presentation.

Corpus Work in a Different Discipline

In discussing the project as a case study of work in a new discipline, I focus on three challenges:

1)*The challenge of compiling a corpus when document types are not always well defined*. Even engineers themselves often cannot consciously describe categories thoroughly. I discuss approaches for categorizing texts from different perspectives, ranging from categories found through linguistic analysis to the perspective that new hires in firms would likely have.

2) *The challenge of working in a discipline with a large distinction between academic and practitioner contexts*. Although I anticipated a gulf between my background in applied linguistics and the knowledge needed for understanding engineering texts, I did not anticipate the skills needed for navigating between academic and practitioner contexts within civil engineering. I discuss ways that this gulf has proven to be a more complicated challenge than the disciplinary difference and the impact it has had on the teaching applications of the research.

3) *The challenge of acquiring and understanding non-academic texts*. For many researchers, including corpus linguists, gaining access to texts outside of academic or research contexts can be daunting. Rarely do studies include collaborators who work in non-academic contexts.

Nevertheless, non-academic contexts are the target for the majority of students in fields such as engineering, so improving students' preparation means we have to understand the practitioner context better.I discuss crucial factors that have allowed me to include the practitioner perspective in this project.

Acknowledgements

Partial support for this project was provided by the United States National Science Foundation's Course, Curriculum, and Laboratory Improvement Program under Award No. 0837776. All opinions, findings, and recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Biber, D. 1988. *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Blakeslee, A. 2001. Bridging the workplace and the academy: Teaching professional genres through classroom-workplace collaborations. Te*chnical Communication Quarterly*, 10(2), 169-192.

Reave, L. 2004. Technical communication instruction in engineering schools. *Journal of Business and Technical Communication*, 18(4), 452-490.

Tenopir, C. and King, D. 2004. *Communication Patterns of Engineers*. Hoboken, New Jersey: IEEE Press/John Wiley & Sons.

Winsor, D. 1996. *Writing Like an Engineer: A Rhetorical Education*. Mahwah, New Jersey: Lawrence Erlbaum.

Multi-word Units and the Academic Word List: Approaches to Research and Pedagogy

Averil J. Coxhead

School of Linguistics and Applied Language Studies
Victoria University of Wellington, New Zealand

Keywords**:** AWL, multi-word units, collocations, lexical bundles, language teaching

The Academic Word List (AWL) (Coxhead, 2000) is now ten years old. This widely known and used list was developed to support English for Academic Purposes students to prepare for study in English-speaking university studies, as well as for independent study. The AWL contains 570 word families. These families are divided into ten sublists, based on the frequency of the whole family. The AWL now appears in a range of textbooks, web pages and applications, mobile phone applications, electronic and paper dictionaries and other teaching materials. It also appears in the Vocabulary Levels Test (Nation, 2001; Schmitt, Schmitt and Clapham, 2001). It has been used as the basis for other word lists such as Coxhead and Hirsh's (2007) pilot science list for English for Academic Purposes (EAP).

Recent research using the AWL has been carried out by a variety of scholars and academic subject areas, including Hyland and Tse (2007), sciences, engineering, and social sciences; Chen and Ge (2007), medical research articles; Konstantakis (2007), Business; Ward (2009), Engineering; Martínez, Beck, and Panza (2009), agricultural sciences research articles; Vongpumivitch et al.(2009), applied linguistics research papers; Li and Qian (2010), Finance; and Coxhead et al. (2010) secondary school science textbooks. This research tends to illustrate that the list covers roughly 10% of academic texts, as Coxhead's initial corpus study showed in 2000. It also shows that the AWL performs slightly lower for sciences than for business or commerce. Debate around the list has ranged from how words in the list behave across different subject areas (see Hyland and Tse, 2007 for example) and the effect of homonyms on the list (Wang and Nation, 2004). See Coxhead (2011) for a discussion of the AWL over the last decade.

Research into different aspects of multi-word units and the AWL is ongoing (Coxhead and Byrd, in press) and not without some challenges. These challenges include working definitions of multi-word units such as collocations and lexical bundles through to variances in the amount of data from words with different frequencies. As we might expect, in some cases, there are strong collocational relationships between words. In other cases, words show strong biases to one side over the other, while some words just seem to like to be alone. Similar patterns of belonging or behaviour occur in the lexical bundles data with AWL words.

A major challenge is how to represent data on multi-word units and the AWL in useful ways for teachers and learners. One approach is to limit multi-word units to specific kinds of words, for example, verbs, nouns, and/or adjectives. Using this approach, it is possible to report on key patterns in the data that might be of interest and help to learners and teachers. A

second approach is more free range. That is, report on every pattern in the data. This more inclusive approach demands a multitude of categories and decisions which may risk presenting the data in overly complex ways. It is important to note also that students' beliefs and approaches to their vocabulary learning can affect learning multi-word units (Coxhead, 2008) and that rationales for supplying particular sets of multi-word data need to be convincing (Byrd and Coxhead, 2010). Furthermore, new research by Boers, Coxhead and Webb (in preparation) suggests that classroom exercises that focus on multi-word units might have a better effect on learning if the exercises present chunks of language holistically, rather than analytically. However, even the chunks themselves do not necessarily give a high return for learning.

These challenges leave us to consider a number of questions. For example, where might research and pedagogy meet over multi-word units and the Academic Word List?What patterns of meaning and use might be hidden by statistical reporting of data?And what principles and considerations might teachers, materials designers, and learners need to keep in mind when considering data on multi-word units?

References

Boers, F., Coxhead, A., and Webb, S.(in preparation). Integral chunks or building blocks: Investigating collocation exercises.

Byrd, P. and Coxhead, A. 2010. *On the other hand*: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL,* 5, 31-64.

Chen, Q. and Ge, C. 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. Eng*lish for Specific Purposes*, 26, 502-514.

Coxhead, A. 2000.A new academic word list. TESOL quarterly, 34 (2), 213-238.

Coxhead, A. 2008. Phraseology and English for Academic Purposes: Challenges and Opportunities. In Meunier, F. and Granger, S. (Eds.), Phraseology in language learning and teaching. Amsterdam: John Benjamins, 149-161.

Coxhead, A. 2011. The Academic Word List ten years on: Research and teaching implications. TESOL Quarterly, 45(2): 355-362.

Coxhead, A. and Byrd, P. (in press). Collocations and Academic Word List: The strong, the weak and the lonely. In Moskowich, I. and Crespo, B. (Eds.) *Encoding the Past, Decoding The Future: Corpora in the 21st Century*. Cambridge: Cambridge University Press.

Coxhead, A. and Hirsh, D. 2007.A pilot science word list for EAP. Revue Française de Linguistique Appliqueé XII (2): 65-78.

Coxhead, A., Stevens, L, and Tinkle, J, 2010. Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics,* 16(2): 35-52.

Hyland, K. and Tse, P. 2007. Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2), 235-253.

Konstantakis, N. 2007. Creating a Business Word List for teaching Business English. *Elia*, *7*, 79-102.

Li, Y. and Qian, D. 2010. Profiling the Academic Word List (AWL) in a financial corpus. Sy*stem*, 38, 402-411

Martínez, I., Beck. S. and Panza, C. 2009. Academic vocabulary in agriculture research articles. Eng*lish for Specific Purposes*, 28, 183-198.

Nation, I.S.P. 2001. Le*arning vocabulary in another language*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55-88.

Vongpumivitch, V., Huang, J. and Chang, Y.2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28(1), 33-41.

Wang, K. and Nation, I.S.P. 2004. Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics,* 25(3), 291-314.

Ethics and Corpus Linguistics

Tony McEnery and Andrew Hardie

Department of Linguistics and English Language
Lancaster University, UK

Keywords**:** corpora, ethics, respondents, legality, replicability

While research ethics are as critical for corpus linguistics as for any other branch of linguistics, relatively little consideration has been paid in the literature to ethical issues in corpus construction and exploitation. Although some authors have directly considered their work in relation to ethical issues, for example Hasund (1998), Sampson (2000) and Rock (2001), the central textbooks in the field, including Sinclair (1991), Kennedy (1998), Biber et al. (1998), and McEnery and Wilson (2001), do not treat ethical issues in any depth. This may be because corpus linguists have in many cases 'inherited' their ethical good practices from guidelines developed, for example, for applied linguistics in general. For example, the British Association of Applied Linguistics has a well-developed set of ethical guidelines which are clearly relevant to corpus builders (see http://www.baal.org.uk/dox/goodpractice_full.pdf).

We will argue, however, that there are questions specific to corpus linguistics which may not be fully addressed by guidelines from outside the field, and thus that research ethics is an area that corpus linguistics should consider in more detail. There are four main groups of such questions. Firstly, in collecting a spoken corpus there are ethical issues relating to the respondents. These relate primarily to privacy – not only of the respondent, but also of the people they are recorded speaking to, and moreover of the people they are recorded talking about. A second set of ethical issues must be addressed in the process of construction of a written corpus. In particular, what is the appropriate attitude to take towards potentially offensive, immoral, or illegal textual data? A third group of questions relate to the sometimes vexed question of the distribution of corpus data. To what extent are corpus distributors ethically obliged to consider whether the purposes to which the data will be put would be approved of by the original donors/collectors of the data? Finally, there are issues that must be faced by any user of corpus data – in particular, the ethical imperatives to take all steps to make sure their analysis is replicable, and to record and preserve aspects of the research method that underlie, but are not contained within, their published results.

While the corpus linguistic literature is mostly silent on ethical issues, it does generally embody good ethical practice. There are, however, a number of exceptions – instances of relatively poor practice in published corpus research. Some occurred during the infancy of corpus linguistics as a (sub-) discipline, but some are more recent. We will review some examples of such poor practice, and suggest that, as a corrective to these relatively prominent bad examples, it is high time that more explicit regard is given to issues of research ethics in corpus linguistics.

References

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics: investigating language structure*

and use. Cambridge: CUP.

Hasund, K. 1998. Protecting the innocent: the issue of informants' anonymity in the COLT corpus. In Renouf, A. (Ed.), *Explorations in Corpus Linguistics*. Rodopi, Amsterdam,13-28.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Harlow: Longman.

McEnery, T. and Wilson, A. 2001. Corpus Linguistics (2nd edition). Edinburgh: EUP.

Rock, F. 2001. Policy and Practice in the Anonymization of Linguistic Data. *International Journal of Corpus Linguistics,* 6(1), 1-26.

Sampson, G.R. 2000. CHRISTINE Corpus, Stage I: Documentation. Available at www.grsampson.net/ChrisDoc.html

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Corpus-based Foreign Language Teaching:

Current Status and Future Directions

Yukio Tono

Graduate School of Global Studies
Tokyo University of Foreign Studies, Japan

Keywords**:** FLT, SLA, learner corpora, CEFR, criterial features

As corpus linguistics is becoming a mainstream methodology within applied linguistics, there is a growing awareness that corpus findings can shed light on various aspects of foreign language teaching (FLT) and second language acquisition (SLA). In this paper, I will first present a methodological framework of corpus-based FLT: the framework of Needs Analysis (Target Situation Analysis and Deficiency Analysis) by Tudor (1997). In addition, three areas of corpus applications in FLT (direct use of corpora, indirect use of corpora and construction of educational corpora), originally proposed by Leech (1997), will be briefly reviewed and integrated into a more comprehensive framework of corpus-based FLT.

Following this framework, I will go on to summarise my past corpus-based research on English language education in Japan in the following fields: (a) the analysis of native speaker (NS) corpora as instantiations of Target Situation Analysis, (b) the analysis of non-native speaker (NNS) corpora as instantiations of Deficiency Analysis, and (c) the analysis of EFL textbook corpora as instantiations of Input Analysis. The findings from NS corpora reveal some significant facts about the role that core vocabulary plays in language use, and how Japanese-speaking learners of English are lacking in productive knowledge of such vocabulary, attested in NNS corpora. It is also found that English textbooks used at Japanese secondary schools have many problems in terms of the coverage of basic core vocabulary and the amount of input. I will also present how such findings were used to design teaching materials in order to bridge the gap between the present and the target situations. My corpus-based TV English conversation series as well as related products based on my TV programs, other teaching materials, self-taught books, various e-learning tools and software will be described as such examples.

In the second part of my talk, I will turn to more recent innovations in corpus-based FLT with a special emphasis on the research into identifying criterial features for the levels proposed by the Common European Framework of Reference for languages (CEFR). By summarising the work by the team at Cambridge, especially the English Profile Programme (EPP henceforth), I will argue that it would be extremely important to identify linguistic properties that are found to be criterial for the given proficiency level. Then I will elaborate on some of the methodological issues of identifying criterial features, focusing on the formal procedures of how to objectively claim that certain features are criterial. To exemplify the case, I will report on my several studies of extracting criterial features from Japanese EFL learner corpora, using semi-automatic error tagging and statistical techniques such as variability-based neighbor coustering and random forest. Also the way to determine the relative values of criterial features as key predictors on CEFR

levels will be discussed, as the number of criterial features increase for different levels. Finally, I will summarise future directions for more integrated approaches for corpus-based FLT.

References

Tudor, I. 1997. *Learner Centredness as Language Education*. Cambridge: Cambridge University Press.

Leech, G. 1997. Teaching and Language Corpora: A convergence. In Wichman, A., Fliegelstone, S., McEnery, T. and Knowles, G. (Eds.), *Teaching and Language Corpora*. London: Longman, 1-22.

A Corpus-Based Investigation into Malaysian ESL Learners' Use of Interactional Metadiscoursein Constructing Their Gender Identity

Roslina Abdul Aziz[1] and Chin Chiu Jin[2]

[1]*Academy of Language Studies*
Universiti Teknologi MARA Pahang, Malaysia

[2]*Institute of International Languages MMU*
Multimedia University, Malaysia

Keywords :ESL, gender identity, metadiscourse, interactional resources, concordance

This paper aims to explore how interactional metadiscourse resources are used to articulate and construct gender identity among ESL learners in Malaysia. Drawing on corpus-based methodology and Hyland's (2004) metadiscourse model as the analysis framework, we perform quantitative and qualitative analyses on the similarities and differences between male and female ESL learners' use of interactional metadiscourse resources, namely hedges, boosters, attitude markers, engagement markers and self mentions (Hyland, 2004, 2005; Tse and Hyland, 2008). The learner corpus used in this study consists of argumentative essays written by third semester ESL learners from diploma and pre-degree programs at two higher learning institutions in Malaysia. The corpus stood at approximately 149 154 words at the time the study was conducted. The corpus was analysed using Antconc 3.2.1, a freeware concordance program developed by Laurence Anthony (2010). The findings from the quantitative analyses show no obvious differences in the average use of attention markers, boosters and hedges between the two genders. However, differences were found in the use of self mentions and engagement markers. The use of these two types of resources by the female learners exceeded the use of the same types of resources by the male learners. The results suggest that there is a relationship between the use of interactional metadiscourse resources and learners' gender, or more specifically, ESL learners in Malaysia actively articulate and construct their gender identity through writing. The findings from this study can provide some valuable insights into the relationship between linguistic variables and non-linguistic variables, particularly gender identity, in shaping ESL learners' second language learning and production.

References

Anthony, L. 2010. AntConc Computer Software (Version 3.2.1 2010) [Computer software]. Retrieved from http: //www.antlab.sci. waseda.ac.jp/.

Hyland, K. 2004. Disciplinary interactions:Metadiscourse in L2 postgraduate writing. Jo*urnal of*

Second Language Writing, 13(2) 133-151.

Hyland, K. 2005. Me*tadiscourse: Exploring interaction in writing*. New York: Continuum.

Tse, P. and Hyland, K. 2008. 'Robot Kung-fu': Gender and professional in biology and philosophy reviews. Jo*urnal of Pragmatics,* 40(7), 1232-1248.

A Corpus-based Investigation into the *be* Overgeneration of *be* + *verb* Constructions by L1-Malay ESL Learners in the Malaysian Corpus of Learner English

Roslina Abdul Aziz[1] and Zuraidah Mohd Don[2]

[1]Academy of Language Studies
Universiti Teknologi MARA Pahang, Malaysia

[2]Department of Languages and Linguistics, University of Malaya
Kuala Lumpur, Malaysia

Keywords: Malaysian Corpus of Learner English (MACLE), ESL, overgeneration, concordance, interlanguage

This study investigates an interesting syntactic problem in the writing of ESL learners whose first language is Malay. It focuses specifically on the overgeneration of *is, are, was* and *were* with inflected and uninflected lexical verbs to form two primary constructions namely;*be +V* and *be + Ved (or Ven* in the case of strong verbs). The study is based on the MACLE ("Malaysian Corpus of Learner English") corpus, which contains 367 argumentative essays and a total of 287 134 words. An analysis of the corpus using Antconc 3.2.1 found uninflected verbs occurred more frequently than inflected verbs in this position after *be.* This seems to confirm the finding of a previous study of Russian learners, in which learners appeared to insert *be* before a lexical verb in their attempts to mark agreement. The *be + Ved/Ven* construction was found to occur with both transitive and intransitive verbs, which suggests that it is not entirely influenced by the class of the lexical verb itself, and could be the result either of learners' attempts to mark tense using inflected forms of *be* or of their inappropriate use of English passive forms. The findings of this study suggest that the overgeneration of constructions with *be* by this group of learners may be the product of a developmental aspect of language acquisition. It can be traced from the system underlying the patterns of overgenerations, which are clearly made up of non-random constructions governed by very specific interlanguage grammar.

References

Anthony, L. 2010. AntConc Computer Software. (Version 3.2.1. 2010), [Computer software]. Retrieved from http://www.antlab.sci.waseda.ac.jp/.

Ionin, T., and Wexler, K. 2001.*L1 Russian children learning English:tense and overgeneration of "be"*. Paper presented at the 2000 Second Language Research Forum.

Contrastive Investigation of Verb-Noun Collocations in the Corpora of ESL Malay andENGLISH Native Learners

Shazila Abdullah[1] and Noorzan Mohd Noor[2]

[1]*Academy of Language Studies*
Universiti Teknologi MARA, Shah Alam, Malaysia

[2]*Academy of Language Studies*
Universiti Teknologi MARA, Shah Alam, Malaysia

Keywords**:** collocations, ESL**,** L2 writing, SLL, collocational competence, learner corpora

A high level of 'collocation', which is commonly described as a string of two or more words that have the tendency to co-occur (Cruse, 1986), is an important aspect in native English writing. This, however, does not seem to be the characteristic of learners' writing. Advanced L2 learners of English may have a wide range of vocabulary but lack the use of effective verb-noun collocations in their writing. Since collocations are formed by combining two words or more, it is imperative to learn them and understand the context in which they can occur. Sadeghi (2009) asserts that without knowing whether or how these words can or cannot occur in certain contexts, one cannot claim to have knowledge of the words since one will have to know the restrictions on co-occurrence of word combinations. A failure to do so may result in ineffective communication since collocations may contribute to learners' communicative competence. Therefore, there is a need for sound knowledge of collocations for ESL learners as the importance of collocations in second language learning is often not realized by them.

As collocation is indeed a difficult feature in L2 writing, it is not surprising that there have been many studies that focus on collocations. Despite that, there can still be many aspects that need reinvestigation. This is especially so in cases where previous studies involved only a relatively small corpus of data and L2 learners with a specific mother tongue. As the influence of mother tongue may not be the same for any particular language, an investigation on Malay ESL learners' writing will thus bridge the gap in the study of collocation in SLL. The emphasis on the importance of learning collocations in second language writing would most probably translate into the need for sound collocational competence for ESL learners, especially at advanced levels, for them to attain both communicative competence and native-like proficiency level in writing.

Verb-noun collocations seem to be one of the most investigated collocations (eg. Nesselhauf, 2003; Koya, 2003; Zinkgraf, 2008; Brashi, 2009). This could be due to the high occurrence of collocations in the academic context and newspaper language that are formed from high-frequency verbs (Biber et al., 1999). Lesniewska (2006: 97) states that "high frequency verbs are very polysemous, the restrictions on their use, which are not predictable from their meaning, may be perceived as highly arbitrary", which would constitute a problem for ESL learners in using them especially the verb-noun collocations. Nevertheless, according to Howarth (1998), it is the ability to use these collocations that will be reflected in the nativeness of their writing. Therefore,

this study will undertake a corpus linguistic approach to explore, identify and contrast the salient features of verb-noun collocations through a contrastive analysis on the linguistic patterns of the common verb-noun collocations used in L1 and L2 learners' writing. Thus, it is the aim of this paper to examine learners' actual use of verb-noun collocations in writing to be able to investigate the extent of their abilities that would reflect on the quality of writing with the hope that it would be able to offer insights on improving the collocational competency of L2 learners in the use of verb-noun collocations with information on the disparities of verb-noun collocations between specific learner corpora being used to provide indications on the problems faced by Malay ESL learners in using lexical verbs and verb-noun collocations.

The study comprises two sets of corpora, i.e. native learner corpus and L2 learner corpus, where quantitative and qualitative analyses are conducted with the aid of a corpus tool. A specific learner corpus was compiled to meet the objectives of this study. Although there are a number of learner corpora for research purposes available in Malaysia, namely, Malaysian Corpus of Learner English (MACLE), Corpus of Malaysian English (COMEL) and the English of Malaysian School Students (EMAS) corpus, the compilation of the present learner corpus is seen as a valuable addition as it consists of the writings of ESL learners from one specific pre-degree programme, i.e. pre-law. These students would have had a considerable high achievement, i.e. a distinction in their English paper in the national examination at the end of their secondary years as it is an admission requirement of the programme. Subsequently, upon completing the pre-degree programme successfully, they would be absorbed into the Law Faculty which requires them to have a high level of English proficiency for them to cope well with the language used in the legal environment, in both the academic context and their future profession. Therefore, the likelihood of using collocations in their course of study and line of work is immense. In addition, the use of a corpus is seen as significant for making generalizations based on the findings of the study.

Acknowledgements

The compilation of the learner corpus used in this study was made possible with the permission given by the Dean of the Academy of Language Studies, UiTM Shah Alam, Malaysia.

References

Biber, D., Johansson, S.,Leech, G.,Conrad, S.and Finegan, E.1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brashi, A. 2009. Collocability as a problem in L2 production. *Reflections on English Language Teaching* 8 (1), 21-34.

Cruse, D.A. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Howarth, P. 1998. The Phraseology of Learners' Academic Writing. In Cowie, A.P. (Ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 161-186.

Koya, T. 2003. A Study of Collocation in English and Japanese Noun-Verb Combinations. *Intercultural Communication Studies* XII (1). 125-147.

Lesniewska, J. 2006. Collocation and second language use. *Studia Linguistica Universitatis Iagellonicae Cracoviensi*s 123, 95-105

Nesselhauf, N. 2003. The Use of Collocations by Advanced Learners of English and some Implications for Teaching. *Applied Linguistics* 24 (2), 223-242

Sadeghi, K. 2009. Collocational Diffrences Between L1 and L2: Implications for EFL Learners and Teachers. *TESL eCanada Journal* 26 (2), 100- 124.

Zinkgraf, M. 2008. V+ N miscollocations in the written production of university level students. *Estudios de linguistica aplicada,* 8, 91-116.

Corpus-based Analyses of Subject and Predicate Concord in Russian

Shinichi Akiyama

Tokyo University of Foreign Studies,
Tokyo, Japan

Keywords**:** Russian, subject, predicate, concord, numeral phrases, multi-regression analysis

Introduction

I have argued on the concord (and non-concord) of subjects and predicates in modern Russian sentences, from a syntactic viewpoint. In particular, I have focused on sentences that contain a numeral or quantifier phrase in the subject. These sentences have two variants of predicates (third person singular vs. third person plural in non-past tense; neuter singular vs. plural in past tense or in predicates that consist of adjectives and particles). See examples (1) and (2) below:

(1) *Работают* *сто* *человек.*
Rabota-ût *st-o* *čelovek.*
work-PRES.3PL hundred-NUM.NOM people-PL.GEN
"There are a hundred people working."

(2) *Работает* *сто* *человек.*
Rabota-et *st-o* *čelovek.*
work-PRES.3SG hundred-NUM.NOM people-PL.GEN
"There are a hundred people working."

I have applied a corpus-based analysis and a statistical method to draw an objective conclusion on this issue. To determine the correlation between categorical distinctiveness (morphology, syntax, etc.) in subjects and subject-predicate concord, I applied a chi-square test. Furthermore, to measure the influence of categorical distinctiveness on subject-predicate concord, I applied multi-regression analysis. The data obtained from these operations can be used to identify which predicate to use in a sentence containing a numeral or quantifier phrase in the subject.

Results

Taking into consideration what has been noted by preceding studies, I list eight categories that may have correlations with subject-predicate concord. These categories include morphological, syntactical, semantic, lexical, and scriptural aspects. Chi-square tests reveal correlations between these categories and the significant preference for singular (SG) or plural (PL) form in the predicates (see Table 1 below).

(Table 1) Correlations between the categories and the significant preference for SG/PL form

morphological	kind of predicates	verbal predicates	SG < PL

		participial predicates	SG > PL
		adjectival predicates	non-significant
syntactical	element order	subject–predicate	SG < PL
		predicate–subject	SG > PL
		sub.1–predicate–sub.2	SG > PL
	existence / non-existence of nominative marker	with PL.NOM. marker	SG < PL
		without NOM. marker	SG > PL
		others	SG < PL
semantic	semantics of numerals[1]	“два” (“two”)	SG < PL
		“три” (“three”)	SG < PL
		“пять” (“five”)	SG > PL
		“сто” (“hundred”)	SG > PL
		“тысяча” (“thousand”)	SG > PL
		“миллион” (“million”)	SG > PL
	animacy of a noun combined withnumerals in subject	animate	SG < PL
		inanimate	SG > PL
	existence / non-existence of approximate number expressions	with approx. num. expressions	SG > PL
		without approx. num. expressions	SG < PL
lexical	“быть” predicate/ non-“быть” predicate	“быть” predicate	SG > PL
		non-“быть” predicate	SG < PL

On the other hand, multi-regression analysis will reveal the diversity of the influences of the categories on subject-predicate coordination, which is expressed by the hierarchy shown below.

(Table 2)The hierarchy of influence of the categories on subject-predicate concord

element order > “быть” predicate/non-“быть” predicate > semantics of numerals > existence/non-existence of nominative marker > animacy of a noun in subject > kind of predicates > existence/non-existence of approximate number expressions

References

Comrie, B. 1989. *Language universals and linguistic typology. Second edition.* University of Chicago Press.

Corbett, G.G. 1978. Universals in the Syntax of Cardinal Numerals. *Lingua,*46(4), 355-368.

Corbett, G.G. 2000. *Number.* Cambridge University Press.

Corbett, G.G. 2006. *Agreement.* Cambridge University Press.

Graudina. L.K.et al. 1976. *Grammatičeskaâ pravil'nost' russkoj reči.* Moscow.

RG1980. *Russkaâ grammatika*. II. Akademiâ Nauk SSSR. Moscow.

Ueda, T. et al. 2003. *Jissen Workshop: Excel tettei-katsuyou Tahenryou-kaiseki.(A Practical Work-shop on Multiple Regression Analyses with Excel.)* Shuwa-systems. Tokyo.

[1] Along with the classification method shown in Corbett 1978.

Identification and Automatic Correction of Common Article Errors in Asian Learner Writing

Laurence Anthony

Center for English Language Education (CELESE)
Faculty of Science and Engineering, Waseda University, Tokyo, Japan

Keywords**:** article errors, error correction, learner writing, natural language processing

Introduction

Asian learners of English are known to struggle with the English article system (*a, an, the*), particularly in their writing (Master, 1987, 1995; Celce-Murcia and Larsen-Freeman, 1999; Han et al., 2006). Previous research has shown that article errors can account for over a quarter of all errors in a given learner corpus (Gamon et al., 2008), and thus, it is no surprise that many books (e.g., Claire and Greenwood, 1988; Brender, 1997; Cole, 2000) and research papers (e.g., Master, 1995; Bitchener, Young and Cameron, 2005; Farrow, 2008; Chuang, 2009) focus on identifying article errors and proposing effective ways to correct them in the classroom. Despite this interest, researchers have shown disagreement about which article errors cause learners the most difficulty. Han et al. (2006: 115), for example, report that the use of the Ø (zero or no) article can be one of the most complex problems faced by non-native speakers of English. On the other hand, Farrow (2008: 30) treats this as less challenging than many other article rules, such as the use of 'the' before superlatives (e.g., 'the nicest guy I ever met') and universally unique items (e.g., 'the sun').

Compounding the problem of article usage is the fact that there are few, if any, mainstream electronic grammar checkers that are able to automatically identify and correct article errors with any degree of accuracy. As part of this study, an investigation of the grammar checker included in Microsoft Word revealed that less than 1% of articles could be identified when they appeared in naturally occurring Japanese learner writing. Others (e.g., Chen, 2009) have investigated the accuracy of automatic grammar checkers designed specifically for second language learners, but again, the results are not encouraging.

This paper addresses two research questions: 1) What are the most common article errors that Asian students commit in their writing? and 2) How can an automatic grammar checking system be developed to automatically identify and correct article errors? To address the first question, I conduct an analysis of article errors found in a corpus of first- and second-year Japanese university student essay reports. In many previous studies on writing errors, highly detailed error tagging systems have been employed, but little attention has been given to the difficult task of identifying what actual errors were committed by the learner. For example, if an article is missing, it is often impossible to identify if the error is an article error or simply an error resulting from the target noun being written in the singular form instead of the correct plural form. To deal with this problem, the error tagging here was carried out by an in-class teacher as part of his regular

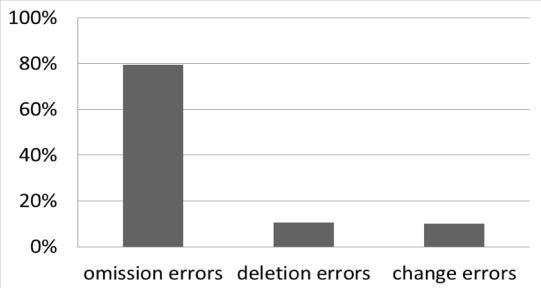


Figure 1. Distribution of Article Errors in a Corpus of Japanese Learner Writing

Table I. Performance measures of proposed system based on a test sample of 50 sentences each containing one error

$Precision_{omission}$	27/27	100%	$Recall_{omission}$	27/38	71%
$Precision_{deletion}$	5/5	100%	$Recall_{deletion}$	5/8	62%
$Precision_{change}$	0/0	-	$Recall_{change}$	0/4	0%
Accuracy = 33/50 = 64%					

teaching duties. As a result, all error tags (and subsequent corrections) were returned to the learner for confirmation. In cases when the articles were mistagged, these were corrected prior to the analysis. Results of this study support those of Han et al. (2006) with omission errors accounting for 79% of all article errors, and deletion errors and incorrect article errors accounting for the remaining 11% and 10% of errors, respectively (see Figure 1).

To address the second question, I explain the design and development of a novel automated article error correction tool. To explain the design of the system, I will first review two common approaches to automatic error analysis systems; 1) data-driven approaches (e.g., Turner and Charniak, 2007), and 2) rule-based approaches (e.g., Bond, Ogura and Ikehara, 1994; Heine, 1998). I will then explain their advantages and disadvantages in terms of their flexibility to be adapted to specialized contexts and domains, their ability to be understood by human users, their scalability, and their speed of processign. In this research, a novel rule-based approach to automatic article error analysis is adopted. Using real-world test data, I will show how the system is simple in design, fast, and accurate, being able to identify and correct the most common learner article errors with 100% precision (see Figure 2). I will also show how the rules adopted by the system can be made accessible to teachers providing them with a useful resource to rank and then teach the most effective article rules to learners.

References

Bitchener, J., Young, S. and Cameron, D. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3), 191-205.

Bond, F., Ogura, K. and Ikehara, S. 1994. Countability and number in Japanese to English machine translation. In Coleman, D. (Ed.), *Proceedings of the 15th Conference on Computational Linguistics. Ky*oto: Association for Computational Linguistics, 32-38.

Brender, A. 1989. *Three little words a, an, the*. Japan: McGraw-Hill.

Celce-Murcia, M. and Larsen-Freeman, D. 1999. *The Grammar Book*. Boston: Heinle & Heinle.

Chen, H. 2009. Evaluating Two Web-based Grammar Checkers -Microsoft ESL Assistant and NTNU Statistical Grammar Checker. *Computational Linguistics and Chinese Language Processing,* 14(2), 161-180.

Chuang, W. 2009. The Effects of Four Different Types of Corrective Feedback on EFL Students' Writing in Taiwan. *Dayeh University Bulletin,* 4, 123-128.

Claire, E. and Greenwood, R. 1988. *Three little words: a, an, and the*. Dundee, IL: Delta Systems Co.

Cole, T. 2000. *The article book: Practice towards mastery of a, an, and the*. Ann Arbor, Michigan: University of Michigan Press.

Farrow, N. K. 2008. Learning to Use the Articles, A and The, in One Lesson. *Asian EFL Journal*, 26, 26-50.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D. et al. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing*. Hyderabad, India: Asian Federation of Natural Language Processing, 449-455.

Han, N.-R., Chodorow, M. and Leacock, C. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2), 115-129.

Heine, J. E. 1998. Definiteness predictions for Japanese noun phrases. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Montreal: Association for Computational Linguistics, 519-525.

Master, P. 1987. A cross-linguistic interlanguage analysis of the acquisition of the English article system. *Unpublished doctoral dissertation*, University of California, Los Angeles.

Master, P. 1995. Consciousness raising and article pedagogy. In Belcher, D. and Braine, G. (Eds.), *Academic writing in a second language: Essays on research and pedagogy*. Norwood, NJ: Ablex Publishing Corporation, 183-204.

Turner, J. and Charniak, E. 2007. Language modeling for determiner selection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, NY: Association for Computational Linguistics, 177-180.

Use of Hedging in Korean EFL Learners' Academic Writing: A Corpus-based Approach

Juhyun Back

Department of English Education

Kyungpook National University, Daegu, South Korea.

Keywords: hedging, learner corpora, English for academic purposes, second language writing, pragmatic transfer, face-saving strategy

The appropriate use of hedged expressions is a particularly problematic feature of non-native writers' English academic writing. Despite its importance, there have not been many studies conducted to investigate how L2 writers can present assertions in their academic writings. In order to investigate how far cultural differences can affect acquiring L2 learners' pragmatic devices, this study compares the features of hedging in NNS and NS academic essays by using two different corpora from Korean masters' students and from English native speakers in the UK. The corpus of target group is taken from 100 Korean EFL university students enrolled at a MA programme and 120 L1 writers' academic essay samples extracted from a corpus of British Academic Written English (BAWE). As a complementary work, the think-aloud procedure was also undertaken with 12 Korean MA sample students to shed light on the corpus data by uncovering the reasoning behind their choices of epistemic devices and their transfer. The use of think-aloud protocols aims to reveal the writer' intentions and to understand 'the thought pattern' of the writer and how far they are related to their patterns in their L1 written discourse (Mastuda 1997:50). The corpus-based analysis focuses on the range and coverage of lexical items as markers of uncertainty and tentativeness, examining the salient features of hedging used in the L2 writings. The results from frequency analysis showed that the Korean student samples show an imbalanced distribution in the use of epistemic devices: they showed an overreliance on a limited set of linguistic items, with the excessive use of modal verbs, but a lower frequency in the use of other lexical devices such as adverbs and lexical verbs.

Based on a detailed contextual analysis of the inventory of epistemic items of Korean EFL students' writings, we examined a range of specific patterns illustrating how the Korean students produce assertions in English academic essays. First, the students showed a lack of use and appropriate manipulation of ways of expressing certainty and doubt in their claims. They tend to present epistemically stronger and direct statements, as they employ modal verbs and adverbs that express the degree of certainty more frequently than the native sample students. The total percentage of *would*, *may*, and *might*, used as probability and possibility

markers in Hyland and Milton (1997)'s category, comprises a lower proportion in the Korean student samples than the native student samples. Conversely, certainty markers such as *will*, *of course*, and *obviously*' shows a high proportion in the Korean student samples. The findings suggest that Korean student samples have a higher incidence of obligatory and authoritative tone in their assertion than that of native speakers. This can be explained by two possible assumptions: the students show the both lack of understanding of the conventions across different text types and linguistic competence in the appropriate use of epistemic devices in written academic prose. From a sociopragmatic perspective, this is partly associated with the differences in rhetorical traditions between L1 and L2 but is not a pragmatic transfer of culturally preferred conventions from L1 to L2. Korean students tend to choose an indirect way of persuading people of their views or ideas in discourse, and thus avoid any confrontational mode of argumentation (Cortazzi and Jin 1997). The results from the think aloud protocols may confirm that their over-generalized perception of the explicitness and directness of English writings may have led to a negative impact on using hedging devices inappropriately. Participant G commented, "I know that presenting assertions in English writings substantially should be clear and direct. However, when I write in Korean, I tend to use linguistic devices that make my claims more politely and indirectly. .... Understanding such difference, I consciously try to avoid indirect and hesitant expressions and thus use such devices like' 'actually', 'indeed', more frequently rather than 'probably' or 'possibly' to make my claim more confidently in English".

Secondly, they seriously lack genre-specific register awareness. They showed a strong preference for using spoken style hedging devices, as the typical features of spoken-like items most frequently repeated in the Korean student corpus exemplified.

(1) **Of course**, it is partly true that homework is ineffective as a learning aid and may actually diminish students' interest in learning. (concordance 16)

(2) However, **personally I think** it is necessary to change the examination system to be able to apply to proper language awareness approach in Korea. (Concordance 15)

(3) **I would like to** talk about universe language awareness, not just about English one this time.

When expressing their own viewpoint, the Korean students have a serious problem of heavily relying on 'talk about' indicating 4.7% of total, which contrasts with the result of 0.45% used by the native student samples. It is, in particular, notable that they tend to heavily depend on personalizing the statements with 'I' as in (2) and (3), often repeated as the forms of 'point-of-view distancing' devices, while few of these appear in the native student samples. This can be partly also explained by the culturally preferred pragmatic transfer from their L1. In fact, several types of politeness strategies, such as expressing personal desire or perspectives, are common in Korean spoken discourse (Koo, 2004)[2]. Korean learners tend to express more or

[2] a. jeoroseonun hyaksilhi malsumdrigo sypsumnida
(I would like to say clearly)
b. iee bun pyungyang sataenun, Jaga bogyenun, ...
(in this event that occur in Pyungyang, in my opinion, ..) (Koo, 2004:6)

less politeness and humbleness toward the reader and simultaneously minimize their responsibility. As participant C in the think-aloud protocols commented, "I found now..in Korean, direct and assertive expressions may cause rudeness toward both listeners and readers. Thus, in Korean, I try to highlight how far I present my personal opinions and views in a polite way. This is very important to keep harmonious mood in interaction with listeners. I found out that this habitually occurs in my English writings". In contrast with direct and authoritative commitments with modal verbs, the Korean samples students tend to use face-saving devices, such as 'point-of-view distancing strategies, in a set of spoken style hedges more often than L1 writers. This contrastive pattern shows their pragmatic failure from L1 to L2 in the two different ways of culture-related transfer and gives some pedagogical implications for teaching alternative strategies to raise both culture-specific and genre-specific awareness in the area of English academic writing.

References

Matsuda, P. K. 1997. Contrastive rhetoric in context: A dynamic model of L2 writing. *Journal of Second Language Writing*, 6 (1), 45-60.

Cortazzi, M. and Jin, L 1997. Communication for learning across cultures. In McNamara, D. and Harris,R. (Eds.), *Overseas students in Higher education: Issues in teaching and learning*. New York: Routledge, 76-90.

Hyland, K. and Milton, J. 1997. Qualification and certainty in L1 and L2 students'' writing. *Journal of Second Language Writing*, 6 (2), 183-205.

Koo, H.J. 2004. A study of aspects of politeness strategy. *Discourse and Cognition*, 11 (3), 1-23.

Corpus Linguistics and The Big Bang Theory

Monika A. Bednarek

Department of Linguistics
University of Sydney, Sydney, Australia

Keywords**:** popular culture, characterization, key words, key clusters, concordances

This paper engages with data from one of the most popular types of texts that we encounter: television dialogue. With a few recent exceptions (e.g. Rey, 2001; Baker ,2005; Mittmann, 2006; Quaglio, 2008, 2009; Bednarek, 2010), corpus linguistics has so far not attended to the language of television series. This is despite the fact that in many countries we spend more hours watching television than consuming other media. In Australia, for instance, the average time spent watching television is more than 3 hours every day (Dick, 2010). Other reasons for studying television dialogue include its influence on viewer's language (Mittmann, 2006) and its function as 'transmodern teacher' (Hartley, 1999), with fictional television series offering models to viewers of how things are done (Wodak, 2009). Indeed, the linguistic study of television dialogue can be seen as an important emerging area of research, with current studies (e.g. Bubel, 2006; Wodak, 2009; Richardson, 2010; Toolan, 2011) drawing on existing research in conversation analysis, (critical) discourse analysis, sociolinguistics, and stylistics. In this paper I aim to show how we can approach the study of television dialogue from a corpus linguistic perspective. I do this through a case study of the sitcom *The Big Bang Theory* (CBS, 2007-present), in particular of the language used by one of the main characters, Sheldon. Using concordances and key words/cluster analysis (Scott ,2004; Scott and Tribble, 2006) I examine how Sheldon is characterised through specific and repeated 'textual cues' (Culpeper, 2001). The analyses clearly show how Sheldon's behaviour indexes his identity as a full-blown nerd, or in Culpeper's (2001) terms, how it 'instantiates' a nerd schema.

References

Baker, P. 2005. *Public Discourses of Gay Men*. London/New York: Routledge.

Bednarek, M. 2010. *The Language of Fictional Television. Drama and Identity*. London/New York: Continuum.

Bubel, C. 2006. *The Linguistic Construction of Character Relations in TV Drama: Doing Friendship in* Sex and the City. Unpublished PhD dissertation, Universität des Saarlandes, Saarbrücken, Germany. Available at http://scidok.sulb.uni-saarland.de/volltexte/2006/598/

Culpeper, J. 2001. *Language and characterisation: People in plays and other texts*. London: Longman.

Dick, T. 2010. Behold the new golden age of TV. *The Sydney Morning Herald News Review*, 23-24 October 2010: 6-7.

Hartley, J. 1999. *Uses of Television*. London/New York: Routledge.

Mittmann, B. 2006. With a little help from *Friends* (and others): Lexico-pragmatic characteristics

of original and dubbed film dialogue. In Houswitschka, C., Knappe, G. and Müller, A. (Eds). *Anglistentag 2005, Bamberg – Proceedings*. Trier: WVT, 573-585.

Quaglio, P. 2008 Television dialogue and natural conversation: Linguistic similarities and functional differences. In Ädel, A. and Reppen, R. (Eds). *Corpora and Discourse. The Challenges of Different Settings*. Amsterdam/Philadelphia: John Benjamins, 189-210.

Quaglio, P. 2009. *Television Dialogue. The Sitcom* Friends *vs. Natural Conversation*. Amsterdam/Philadelphia: John Benjamins.

Rey, J. M. 2001 Changing gender roles in popular culture: Dialogue in *Star Trek* episodes from 1966 to 1993. In Biber, D. and Conrad, S. (Eds). *Variation in English: Multi-dimensional studies*. London: Longman, 138-156.

Richardson, K. 2010 *Television Dramatic Dialogue. A Sociolinguistic Study*. Oxford: Oxford University Press.

Scott, M. 2004. *Wordsmith Tools.* Version 4. Oxford: Oxford University Press.

Scott, M. and Tribble, C. 2006. *Textual patterns. Key words and corpus analysis in language education*. Amsterdam/Philadelphia: Benjamins.

Toolan, Michael. 2011. 'I don't know what they're saying half the time, but I'm hooked on the series': incomprehensible dialogue and integrated multimodal characterisation in *The Wire*. In Piazza, R., Bednarek, M. and Rossi, F. (Eds). *Telecinematic Discourse: Approaches to Language in Fictional Cinema and Television*. Amsterdam/Philadelphia: John Benjamins, 161-183.

Wodak, R. 2009. *The Discourse of Politics in Action*. Basingstoke/New York: Palgrave Macmillan.

Error Tagging a Malaysian Learner Corpus: Pitfalls and Rewards

Simon Philip Botley
Academy of Language Studies
MARA University of Technology (UiTM), Kota Samarahan, Sarawak, Malaysia

Keywords**:** corpus, learner corpus, CALES, error tagging, Computer-aided Error Analysis

Introduction

Teachers of English as a Second Language (ESL) in Malaysia have a high level of awareness of the mistakes or errors made by their students in class. However despite this awareness, it can be said that not all ESL educators can or do use this information in order to help students to avoid making errors in the first place. This is partly because up to fairly recently, there has been a lack of reliable and permanent data on learner performance in English among Malaysian students. Such data, if it was available, could be used as to help ESL educators to predict what kind of errors students make, and to do something about the errors in a pedagogically plausible manner.

At the moment, most ESL educators in Malaysia seem to rely upon their professional experience or linguistic intuitions to predict what kinds of errors in the L2[3] will be made by a particular L1[4] group. For instance, it is widely known among Malaysian ESL educators that Malaysian learners of English regularly over-use the definite article (*we must work for the money* versus *we must work for money*), and turn non-countable nouns onto countable ones (*a staff*, rather than *a member of staff*).

Errors such as these may be traced back to the L1 which in most cases is Bahasa Malaysia (Malay), a language which does not have a system of definite and indefinite articles, and in which the notion of countability is somewhat different to that in English.

Despite the fact that we already have some notion of which error features are likely to be encountered in students' written performance, what is missing, at least in Sarawak, is a permanent collection of large amounts of student 'error data', taken from different L1 backgrounds, stored in a machine-readable form and available for pedagogical and research purposes. This is the reason why the CALES learner corpus was created.

Objectives

This paper discusses some practical and theoretical issues surrounding Computer-aided Error Analysis (Granger et al., 2002, Granger, 2004), in particular the (semi-) manual error tagging of a new learner corpus in Malaysia. This corpus, called CALES (Corpus-based Archive of Learner English in Sabah/Sarawak) (Botley et. al, 2005, 2007; Botley and Metom, 2011), contains

[3] Second or target language
[4] First or native language

argumentative essays written by students at a number of universities in East Malaysia. The size of the corpus currently stands at just over 480,000 words.

Methodology and Results

The data was collected according to the International Corpus of Learner English methodological guidelines (Granger, Meunier, and Dagneaux, 2002). Essays were written by students in class and later digitized. Information about the students was recorded on learner profile instruments, to provide independent variables for comparison and contrast.

A 32,000 word sample of the corpus was manually annotated using the Universite Catholique du Louvain Error Editor and its tagset (Dagneaux et al, 1998). The tagset consists of a fairly delicate system of alphanumeric tags that encode a wide range of primarily lexical and grammatical error categories. Once tagged, the texts were run through the Wordsmith Tools concordance software (Scott, 2008) to produce error frequency statistics, so as to provide empirical findings on frequent and rare error categories in the data. The analysis also included a comparison of the frequency counts from essays written by students from two major ethnic groups in Sarawak, namely Malays and Ibans.

As well as providing some empirical findings, the main contribution of this paper is to discuss the various challenges and benefits arising from such work, and how the challenges were overcome, or not. The message of this paper is that despite some practical and theoretical concerns about error tagging, as well as the hard manual labour involved in tagging a large learner corpus, the enterprise pays great dividends in terms of the abundant amount of rich data that can be obtained from such analysis. In particular, this kind of work can reveal many detailed insights into the linguistic features of learners' interlanguage (Selinker 1972) where different groups of learners may come from contrasting L1 backgrounds or ethnicities.

References

Botley, S. P., De Alwis, C., Izza, I. and Metom, L. 2005. *CALES: A Corpus-Based Archive of Learner English in Sarawak*. Final Project Report, Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Botley, S. P., Dillah, D. and Metom, L. 2007. *Corpus Based Archive of Learner English in Sarawak and Sabah (CALES Phase 2)*, Final Project Report, Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Botley, S. P. and Metom, L. 2011. *A Corpus-Based Investigation of the Interlanguage of University Students in East Malaysia. Fi*nal Project Report, Research Management Institute, Universiti Teknologi MARA.

Dagneaux, E., Denness, S. and Granger, S. 1998. Computer-aided error analysis. Sy*stem. An International Journal of Educational Technology and Applied Linguistics*, 26(2), 163-174.

Granger, S. (Ed.) 1998. *Learner English on Computer*. Harlow: Addison-Wesley Longman.

Granger, S., Hung, J. and Petch-Tyson, S. 2002. (Eds.)*Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S., Meunier, F., & Dagneaux, E. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. 2004. Computer learner corpus research: current status and future prospects. In Connor, U. and Upton, T. (Eds.), *Applied corpus linguistics: a multidimensional perspective. Am*sterdam & Atlanta: Rodopi, 123-45.

Scott, M. 2008. *Wordsmith Tools*. Oxford: Oxford University Press.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics,* 10, 209-31.

He Pātaka Kōrero:

Using Corpora of Māori to Inform Resources for Learning and Using Māori in Education and Workplace Settings.

Mary T. Boyce

Department of Indo-Pacific Languages and Literatures
University of Hawai'i at Mānoa, Honolulu, Hawai'i, USA

Keywords: Māori corpora, Māori dictionaries, immersion education resources, Māori legal language

Overview

Sizeable computerized corpora of Māori have been being designed and compiled for more than a decade now. Three of these corpora, the Māori Broadcast Corpus, the Legal Māori Corpus and the Corpus of Māori Texts for Children have been used to inform the production of resources for use in the education and legal domains. The nature and purpose of these three corpora, and how they have been used for resource production will be described. Future work using these corpora and priorities for further corpus development in these domains will be outlined.

Background: early work with smaller corpora of Māori

Early collections of digitized Māori text typically involved the collection of smaller corpora for specific purposes. This work includes:

1. Margaret Mutu's (Mutu-Grigg, 1982) work on Māori manner particles, based on a collection of written texts and responses to oral questions from native speaker informants.
2. Richard Benton (1982) with Hiria Tumoana and Andrew Robb collected a small corpus of texts to use in identifying high frequency words of Māori, and published the first such lists for Māori.
3. Ray Harlow's word list from texts collected by the Reverend J. Watkin in the 1840s,*A Word-list of South Island Māori* (1985), and, the name and word indexes of the traditional prose accounts in *Ngā Mahi a ngā Tūpuna* (1986, with Agathe Thornton), and the song text collections in *Ngā Mōteatea* (1990).

More recent corpora

The more recent corpora of Māori are larger and broader in scope. There are two designed, representative corpora of Māori that are publicly available:

• the Māori Broadcast Corpus (Boyce, 2006), one million words of spoken Māori transcribed from radio and television broadcasts in the mid 1990s. This is the first sizeable corpus of spoken Māori and was designed and compiled to identify high frequency vocabulary in Māori.

- the Legal Māori Corpus (Stephens, Boyce et al., 2010), eight million words of written texts from the legal domain spanning 1828-2009. This is the largest known corpus of Māori to be publicly available (The pre-1910 texts are available, the post-1910 texts are undergoing copyright and confidentiality clearance, and will be released when this is completed). The Legal Māori Corpus was designed and compiled to inform the production of the Legal Māori Dictionary (due 2012), and to provide a pool of data to investigate legal language in Māori. A further outcome of this project is the Legal Māori Archive, also available to the public through the New Zealand Electronic Text Centre at Victoria University of Wellington. (http://www.nzetc.org/tm/scholarly/tei-corpus-legalMaori.html)

There are other corpora of Māori that are not publicly available; these include:

- the Corpus of Māori Texts for Children (Boyce, M. with Huia Publishers 2001 – ongoing) compiled to inform work on *Tirohia Kimihia*, a monolingual dictionary for young learners in immersion education. The Corpus of Māori Texts for Children is currently at approximately 750,000 words. We anticipate work on additional design, balance and compilation of this corpus to resume shortly.
- the large corpus compiled by Te Taura Whiri i te Reo Māori (the Māori Language Commission) while they were compiling *He Pātaka Kupu* (2008), a comprehensive monolingual dictionary for fluent adult users.

References

Benton, R. H. Tumoana and A. Robb 1982. *Ko ngā Kupu Pū Noa o te Reo Māori: the first basic Māori word list.* Wellington: New Zealand Council for Educational Research.

Boyce, M. 2006. *A corpus of modern spoken Māori.* Unpublished PhD thesis, Victoria University of Wellington.

Boyce, M. 2006. The *Māori Broadcast Corpus.* (the corpus files and an electronic copy of the PhD dissertation document are available from Mary Boyce: mboyce@hawaii.edu)

Boyce, M. and Stephens, M. 2010. *The Legal Māori Project.* http://www.victoria.ac.nz/law/research/research-projects/legal-maori/default.aspx

Harlow, R. 1990. *A name and word Index to Nga mahi a nga tupuna.* Dunedin: University of Otago Press.

Harlow, R. and Thorton, A.1986. *A name and word Index to nga moteatea.* Dunedin: University of Otago Press.

Harlow, R. 1985.*A Word-list of South Island Māori.* Auckland: Linguistic Society of New Zealand.

Huia Te Manu Tuku Kōrero. 1986. *Tirohia Kimihia: He Kete Wherawhera. We*llington: Huia Publishers, for the Ministry of Education.

Mutu-Grigg, M. 1982. The *manner particles rawa, tonu, noa, kee and kau in Māori.* MPhil Thesis, University of Auckland.

Stephens, M., Boyce, M.et al. *The Legal Māori Corpus*. (available from:

http://www.victoria.ac.nz/law/research/research-projects/legal-maori/corpus.aspx

Te Taura Whiri i te Reo Māori. 2008. He *Pātaka Kupu: te kai a te rangatira. Au*ckland, NZ: Raupo.

Compiling and Analysing a Spoken Academic Corpus

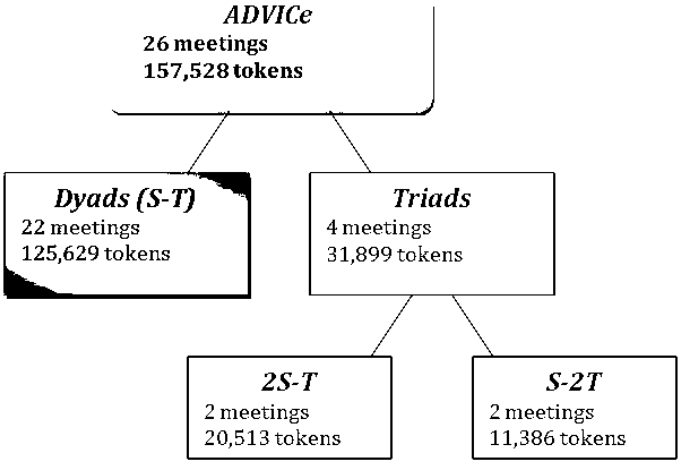
Vaclav Brezina
Department of Applied Language Studies and Linguistics
The University of Auckland, New Zealand

Keywords**:** corpus linguistics, EAP, academic speech, corpus design, corpus tagging, *ADVICe*

Spoken academic English has been generally recognised as a specific genre/register with its own social dynamic and unique linguistic features (Biber, 2006; Swales, 2006). Swales (2006) characterises academic speech as a genre, which is much closer to informal conversation than written academic prose, a genre which contains "a considerable amount of technical lexis embedded into a loosely co-ordinated sentence structure and surrounded by heavy employment of deictic elements" (p. 23). The best way to investigate the typical characteristics of academic speech is through analysis of a corpus.

Building a corpus of academic speech, however, presents a major challenge (Crowdy, 1993, 1994, 1995; Biber et al., 2001; Simpson-Vlach & Leicher, 2006). Unlike samples of academic writing which are available in copious quantities online, academic speech needs to be painstakingly recorded as well as carefully transcribed and annotated.

This workshop draws on the author's experience with building a corpus of academic speech in the New Zealand context. The project has been carried out at the department of Applied Language Studies and Linguistics, University of Auckland (UoA). The result of the project is a small (160,000 tokens) single-genre corpus of spoken university English – *Academic discourse verbal interactions corpus (ADVICe)* – see Figure 1. *ADVICe*, which will be made available to a larger linguistic community, consists of transcribed and morphologically annotated recordings of advisory sessions (office hours, supervisory meetings etc.) between students and lecturers at UoA.



S...student, T...lecturer

Figure1: ADVICe – structure

The workshop focuses on the main issues of spoken corpus design and analysis and compares *ADVICe* with other available corpora of academic speech (e.g. *MICASE* and *BASE*). This will be of particular interest to researchers and practitioners interested in spoken academic discourse in general as well as in particular lexico-grammatical patterns in academic speech.
The following areas will be covered:

- *ADVICe* and other corpora of academic speech (*MICASE, BASE*)
- Data gathering (participants recruitment & selection, recording techniques)
- Transcription (transcription conventions, software options)
- Corpus creation (checking for errors, tagging, storage)
- Corpus analysis

References

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., Reppen, R., Clark, V., & Walter, J. 2001. Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus. In R. C. Simpson & J. M. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 48-57). Ann Arbor: University of Michigan Press.

Crowdy, S. 1993. Spoken corpus design. *Literary and linguistic computing, 8*(4), 259-265.

Crowdy, S. 1994. Spoken corpus transcription. *Literary and linguistic computing, 9*(1), 25-28.

Crowdy, S. 1995. The BNC spoken corpus. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer* (pp. 224-234). Harlow: Longman.

Simpson-Vlach, R. C., & Leicher, S. (2006). *The MICASE Handbook: A Resource for Users of The Michigan Corpus of Academic Spoken English*. Ann Arbor: University of Michigan Press.

Swales, J. M. 2006. Corpus linguistics and English for academic purposes *Information technology in languages for specific purposes* (pp. 19-33). New York: Springer.

Significant or Random? A Critical Review of Statistical Analyses in Corpus-Based Sociolinguistic Studies

Vaclav Brezina[1] and Miriam Meyerhoff[1]
[1]*Department of Applied Language Studies and Linguistics*
The University of Auckland, New Zealand

Keywords**:** sociolinguistics, corpus linguistics, statistical procedures, social variation, meaningful variation

Introduction

This paper offers a critical review of a methodology often employed in corpus-based sociolinguistic studies (e.g. Macaulay, 2002a, 2002b; Xiao and Tao, 2007; Barbieri, 2008; Murphy, 2010) which make use of aggregate data. This methodology relies on a general comparison of frequencies of a target linguistic variable in socially defined sub-corpora (e.g. speech of all men vs. speech of all women in the corpus). An issue with this procedure lies in the fact that it emphasises the inter-group differences and ignores within group variation. The methodology thus often yields falsely positive results (with highly significant log-likelihood scores). The paper presents evidence which shows that sociolinguistic studies based on aggregate data are in principle unreliable. We will demonstrate that random (and therefore sociolinguistically irrelevant) speaker groupings can often yield statistically significant results.

Data and methodology

The analyses are based on *BNC32*, a one-million-word corpus of British informal conversation, which was extracted from the demographic part of the *British National Corpus* (*BNC*). It represents the speech of 32 British English speakers – 16 women and 16 men. The speakers form a balanced sample in terms of gender, age and socioeconomic status.

The following table (Table 1) summarises the main features of the corpus. The major advantage of *BNC32* is the fact that (unlike the majority of commonly used corpora) it enables us to search for language forms in the speech of individual speakers.

Table 1. Structure of BNC 32

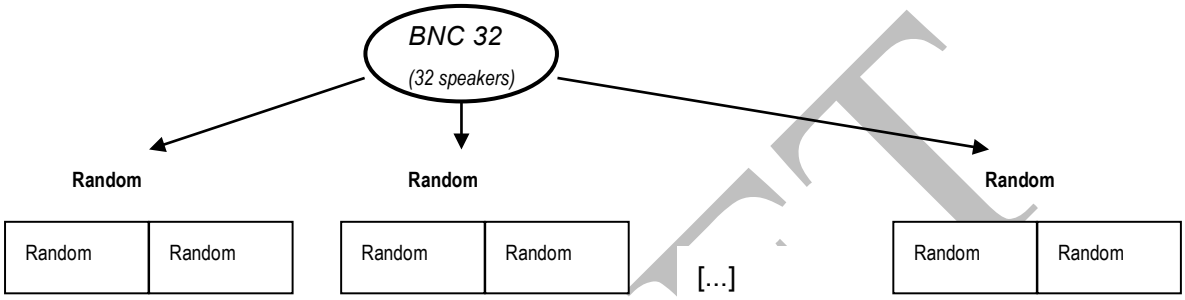
Corpora	Tokens (running words)	No. of speakers	Speaker's gender	Speaker's age	Speaker's socio-economic status	Genre	Discourse mode	Variety of English	Period
BNC 32	1.04 mil.	32	16 male 16 female	15-34: 10 35-54: 13 55+: 9	AB: 7 C1: 9 C2: 9 DE: 5 unknown: 2	informal conversation	highly interactive	UK	early 1990s

The first step of the analysis was to compare the occurrence of seven linguistic forms in *BNC 32* subcorpora based on gender, age and socio-economic status. This comparison was based on

aggregate data with the use of log-likelihood statistic, which is a common procedure in corpus comparison studies (Rayson et al., 2004).

The second step was to apply the same methodology to random speaker groupings. *Random Integer Set Generator* (www.random.org) was used to randomly assign the 32 speakers into 500 group pairs (see Figure 1). This sampling procedure was repeated three times with the final number of 1,500 random assignments. For every pair of random speaker groups, the log-likelihood score was calculated comparing the occurrences of seven dependent variables in each of the 16-speaker groups. Finally, the percentage of statistically significant differences between the pairs of random groups was calculated for each of the dependent variables.

Figure1. BNC 32 - Random grouping



Findings

Consistent with many previous sociolinguistic studies, the results based on aggregate data show an effect for all social variables in question (speaker's age, gender and sociolinguistic status). At the same time, however, a large proportion of randomly created groups of speakers differ significantly from other random groups. Table 2 below presents an overview of percentages of statistically significant results in the three 500 random groupings for seven dependent variables. We can see that with frequent linguistic variables such as the classical hedge *sort of*, the epistemic phrase *you know* or the definite article almost 80 per cent of comparisons of random speaker groupings show statistically significant differences.

Table 2. Random variation in BNC 32: per cent of statistically significant results (p<.05)

Grouping	*kind of*	*sort of*	*the*	*I @think you*	*I @ think I*	*you know*	*ain't*
random 500 I	45.4%	81.4%	80.4%	34.8%	8%	76.2%	76%
random 500 II	44.4%	78.2%	79.2%	40.8%	5.6%	75.6%	74.8%
random 500 III	43.4%	77.8%	79.8%	37.6%	4.8%	82.2%	75.2%
MEAN	44.4%	79.1%	79.8%	37.73%	6.13%	78.00%	75.33%

@...any one or zero words to allow for phrases such as *I don't think I, I just think I, I do think I* etc.

References

Barbieri, F. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1), 58-88.

Macaulay, R. 2002a. Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. *Journal of Sociolinguistics*, 6(3), 398-417.

Macaulay, R. 2002b. You know, it depends. *Journal of Pragmatics*, 34(6), 749-767.

Murphy, B. 2010. *Corpus and Sociolinguistics: Investigating age and gender in female talk*. Amsterdam: John Benjamins.

Rayson, P., Berridge, D., and Francis, B. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In Purnelle, G., Fairon, C.and Dister, A. (Eds.), *Le poids des mots*. Louvain-la-Neuve: Presses universitaires de Louvain,926-936.

Xiao, R., and Tao, H. 2007. A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies*, 1(2), 241-273.

Māori Loanwords in New Zealand English

Andreea S. Calude and Mark Pagel

School of Biological Sciences
University of Reading, Reading, UK

Keywords: Māori, New Zealand English, lexical borrowing, media language, diachronic corpus

Introduction

While many countries around the world are legislating against the flow of borrowed words entering their languages (e.g., France, Germany, Brazil), one country opens its arms to this: New Zealand. New Zealand English is not threatened by Māori loans because ever since the 1987 *Māori Language Act*, New Zealand has been concerned with preserving and promoting the Māori language, which it sees as forming an integral part of New Zealand culture and identity. New Zealanders are encouraged to use Māori anytime, even if only to correctly pronounce their local Māori town.

Data and Methodology

Previous studies (Macalister, 2006, 2007; Kennedy, 2001) suggest that the number of Māori loanwords used in New Zealand English is indeed increasing. However, this data is at least 10 (in case of the Wellington Corpus of Spoken New Zealand English, Holmes et al., 1998) to 15 years old (in case of the Wellington Corpus of Written New Zealand English, Bauer 1993) or more (*School Journals*). The current paper investigates more recent numbers of Māori loanwords by putting together and analysing a chronological (diachronic) corpus of New Zealand newspapers, namely the New Zealand Press Corpus (NZPC). The NZPC consists of samples from New Zealand newspapers from the North and the South Island (*The New Zealand Herald, The Press, The Dominion Post* and *The Southland Times*), taken at 3-year intervals (1996, 1999, 2002, 2005, 2008, 2011), to follow on from the Wellington Corpus of Written New Zealand English.

Findings

The 5 million-word corpus indicates that contrary to existing assumptions and predictions, the number of Māori loans is not increasing, and most loans are in fact decreasing in use. Findings are contrasted and compared across the different newspapers, regions and years. This work has implications for the field of language revitalisation and language planning, and makes contributions to research on borrowing and language contact as New Zealand presents an unusual contact situation: the borrowing of words from a sub-dominant, indigenous language (Māori) into a dominant language (English).

Acknowledgements

AC would like to thank NZFRST (New Zealand Foundation for Research, Science and Technology) for their generous support and MP would like to thank the Research Council Advanced Investigator Grant.

References

Bauer, L 1993. *Manual of Information to accompany the Wellington Corpus of Written New Zealand English.* Wellington: Department of Linguistics, Victoria University of Wellington.

Holmes, J., Vine, B., and Johnson, G. 1998. *The Wellington Corpus of Spoken New Zealand English: a User's guide*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Kennedy, G. 2001. Lexical Borrowing from Māori in New Zealand English. In B. Moore (Ed), *Who's Centric Now? The Present State of Post-Colonial Englishes.* Oxford: Oxford University Press, 59-81.

Macalister, J. 2006. The Māori presence in the New Zealand English lexicon, 1850-2000. *English World-Wide*, 27 (1), 1-24.

Macalister, J. 2007. Weka or woodhen? Nativization through lexical choice in New Zealand English. *World Englishes*, 26 (4), 492-506.

Prosodic Development in the Interlanguage Corpus of Thai Learners

Chun-Mei Chen

Department of Foreign Languages and Literatures
National Chung Hsing University, Taiwan

Keywords**:** prosodic development, CSL, interlanguage, Thai learners

Introduction

This paper investigates how prosodic development of Thai learners of Chinese as a second language (CSL) is revealed in their interlanguage, and how mean length of utterance is correlated with the accuracy of prosodic features in the interlanguage corpus. Corpora have been used for language teaching and learning (Coxhead, 2002, Sinclair, 2003, Römer, 2004), but few empirical studies investigate the effect of corpora on the learning process and second language development. Corpus-based studies primarily have focused on grammar, pragmatic and discourse features (Möllering, 2001, Belz, 2008), and suprasegmental prosodic issues in learners' corpus are seldom explored.

Various kinds of measures have been used to compare the trajectories of child language development. The Phonological Mean Length of Utterance (PMLU) measure utilizes spontaneous speech samples and draws attention from the segments to the word as a whole (Ingram 2002, Saaristo-Helin 2009). The assessment tool of the PMLU method measures whole-word complexity for both child and target words, the degree of accuracy in producing the words. In the context of second language learning, a proficient second language speaker is assumed to be able to perform tasks fluently and accurately. Fluency and accuracy are two aspects of language production with reference to oral task performance of L2 learners. Many language researchers (Lennon, 1990, Towell et al., 1996) analyzed oral production data to determine which quantifiable linguistic phenomena contribute to perceptions of fluency in L2 speech. Accuracy is the ability to avoid error in performance (Skehan and Foster, 1999). Both fluency and accuracy have proved useful measures of L2 performance. However, longitudinal studies on phonological development with second language learners are rare.

In the present study, corpus data were drawn from both laboratory recordings and Mandarin language classrooms. The factors of elicitation materials, topics for conversations, and the familiarity with the production tasks were controlled. Forms and functions of the prosodic elements such as Mandarin pitch contours, durations, pauses, and learners' progress of accuracy were analyzed to justify the effects of mother tongue and length of utterance on the prosodic development of the second language learners.

Methods

The research was carried out on an interlanguage corpus created during the Fall 2010 semester and the Spring 2011 semester. The corpus consisted of 108 hours of video-taping of Elementary

Chinese classroom discourse and voice tokens collected from laboratory recordings in Week 4, Week 8, Week 12, Week 16, Week 20, and Week 24. Twelve second language learners of Mandarin Chinese from Thailand participated in the study. Pitch height and peaks were marked in the corpus transcription, with indication of pauses and durations in the learners' interlanguage.

All the subjects from Thailand had no background of the Chinese language before they enrolled in the language course, and they were either exchange or international students in Taiwan. The target language in the classrooms is Mandarin. Subjects from the same country also used their mother tongue to communicative with each other. The second language discourse was collected from multicultural classrooms

Results

Measurements from the interlanguage tokens have shown the mean length of utterance is correlated with the accuracy of prosodic features. The longer the utterance unit, the more accurate the prosodic features. Pitch range of deviant tokens produced by Thai learners was narrower and flat, and the durations of pauses were longer. The prosodic features of the repeated tokens were also analyzed in the interlanguage corpus. When Thai learners confirmed the target utterance for the first time, the target token has shown the highest pitch with the longest duration among the repetition forms.

On the other hand, Thai learners made salient progress during the first semester of Mandarin learning, from 35% to 87% of correct tonal identification rate in laboratory perception tests from the pretest to Week 12. Variability was also attested among the Thai learners, as some elementary learners have demonstrated better tonal perception rate than the learners enrolled in intermediate-level language courses. Acoustic measurements and discourse analysis of the interlanguage have revealed that silence and repairs are most important cues for the non-target-like utterance unit.

Discussion

This study has argued for the importance of integrating an analysis of corpus-based prosodic measures as a means for second language evaluation. Corpus-based research has an increasing influence on language teaching pedagogy and second language learning. The interlanguage corpus of Thai learners in the present study has provided individualization of the analysis for the purpose of assessment of the variation between learners. Prosodic marking in learner corpora could facilitate the analysis of interlanguage discourse. The investigation on the prosodic development of second language learners of Mandarin Chinese accompanied with the analyses of forms and functions in the interlanguage corpus. Visual prosodic marking in the learner's corpus was presented to the learners in follow-up perception tasks. The integration of prosodic analyses in the learner's corpus facilitated the tonal learning of the Thai learners and linked to the learning process with specific reference to second language learners' oral communicative competence.

References

Belz, J. A. 2008. The role of computer mediation in the instruction and development of L2 pragmatic competence. *Annual Review of Applied Linguistics*, 27, 45-75.

Coxhead, A. 2002. The academic wordlist: A corpus-based word list for academic purposes. In Kettemann, B. and Marko, G. (Eds.), *Teaching and Learning by Doing Corpus Analysis*.

Amsterdam: Rodopi, 73-89.

Ingram, D. 2002. The measurement of whole-word productions. *Journal of Child Language*, 29, 713-733.

Lennon, P. 1990. Investigating fluency in EFL: A quantative approach. *Language Learning*, 40, 387-417.

Möllering, M. 2001. Teaching German modal particles: A corpus-based approach. *Language Learning and Technology*, 5(3), 130-151.

Römer, U. 2004. A corpus-driven approach to modal auxiliaries and their didactics. In Sinclair, J. (Ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 185-199.

Saaristo-Helin, K. 2009. Measuring phonological development: A follow-up study of five children acquiring Finnish. *Language and Speech*, 52(1), 55-77.

Sinclair, J. 2003. *Reading Concordances: An Introduction.* New York: Longman.

Skehan, P. and P. Foster. 1999. The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49 (1), 93-120.

Towell, R., Hawkins, R. and Bazergui, N. 1996. The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84-115.

Developing an EAP Website Supported by Free Academic Spoken Corpora (FASE)

Howard Chen and Shu-Hua Fang

National Taiwan Normal University and National Tsing-Hua University

Keywords**:** academic English corpus, web, spoken English, FASE

As English is the most widely used language in various international academic communities, graduate students in many ESL/EFL countries need to develop good English academic listening and speaking skills. Corpora can play an important role in helping students improve their command of academic English (O'Keeffe and McCarthy, 2007). Information extracted from various authentic native corpora can be used to develop a variety of EAP teaching and learning materials. Two very good examples of academic spoken corpora are MICASE (The Michigan Corpus of Academic Spoken English) and BASE (British Academic Spoken English Corpus). Although these academic spoken corpora have already made significant contributions in language teaching and learning, they remain insufficient in size when compared with larger written corpora such as BNC or COCA.

Although the need for a larger academic spoken corpus is obvious; it is, however, difficult and very time-consuming to record and transcribe spoken data. More recently, however, a new type of language learning resource has emerged. Several top-notch universities such as MIT, Stanford, and Yale have begun to put their high-quality academic courses online. Some of these online courses provide video, audio, and even complete lecture transcripts. These academic resources are valuable, but they are mainly designed for English native speakers. It is thus very difficult for non-native speakers to follow the academic lectures.

To allow ESL/EFL learners and even EAP teachers to make better use of the learning content, we first collected various free academic lectures and then developed a corpus-based English-for-Academic-Purpose website. The academic spoken corpus called Free Academic Spoken English (FASE) now has collected more than 15 million words in academic lectures. This EAP website, which is based on a powerful open source software called The IMS Open Corpus Workbench, has the following features:

1. The academic lectures in different academic disciplines are compiled and POS-tagged, allowing ESL/EFL learners to search any word, phrase, and collocations in FASE. This can help learners to better understand the proper contexts of using various academic words and phrases.
2. The importance of lexical bundles is widely accepted various applied linguists. The lexical bundles extracted from FASE are provided (ranging from trigrams to 5-grams). In addition, academic lexical bundles with similar functions are also grouped together. The

classifications might further help EFL learners properly use the “academic formulas” we provide.

3. To help ESL/EFL learners better comprehend the lecture content, the keyword and key phrases of FASE was also extracted with the help of corpus processing tools. Students will be able to better understand the special features of academic spoken languages via the analyses of the keywords and key phrases.
4. Some audio/video clips are also linked to transcripts with the help of speech recognition technologies. The synchronized video and audio can help learners improve their listening and pronunciation/speaking. When learners click on any words or sentences, they can hear specific words and sentences pronounced. The instant access to various types of input can facilitate their listening comprehension and vocabulary learning.

Developing a Chinese as Second Language Learner Corpus and a Web-based Concordancer

Howard Chen and Christine Yang

National Taiwan Normal University

Keywords**:** learner corpus, CSL, web, concordancer

Many researchers and language teachers around the world believe that language corpora have great potentials for improving second/foreign language learning and teaching. Among various types of corpora resources, learner corpora in particular have received much attention recently. One of the most influential projects is the ICLE (International Corpus of Learner English) project led by Professor Granger in University of Louvain, Belgium. The new version of ICLE corpus contains 3.7 million words of EFL writing from learners representing 16 mother tongue backgrounds. The ICLE corpus has helped to produce many research paper and pedagogical materials within the past decade. Although useful English learner corpora such as ICLE are widely available, very few learner corpora for other target languages are available.

Recently, because of the rapid economic growth in China, an increasing number of students are learning Chinese as a second language. Although the number of
Chinese as second language (CSL) learners is increasing rapidly around the world; very few CSL learner corpora are available for teaching, learning, and research. For
CSL research, learner corpus can play an important role. CSL teachers and researchers can conduct research on learners' interlanguage development and gain insights about leaner's difficulties and needs. Material writers can further use the results of error analysis to produce pedagogical materials. CSL leaner corpus might also be used to better understand the differences among learners at different proficiency levels (cf. Cambridge English Profile Project).

This paper will introduce a new Chinese as second language learner corpus and related corpus search tools developed by MTC (Mandarin Teaching Center) and SC-TOP (Steering Committee of Test of Proficiency) in Taiwan. MTC is located at National Taiwan Normal University and it is the largest Chinese teaching centers in Taiwan. There are more than 1600 students enrolled in each quarter, and there are more than 150 teachers in this center. Students from more than 70 countries are studying in this center. SC-TOP is a language testing research center sponsored by Ministry of Education for developing various Chinese as a second language proficiency tests. Based on the data provided by these two centers, a 3.5 million words Chinese as a second language learner corpus has been developed. The MTC-

TOP leaner corpus includes the following three different types of learner data-
1.CSL learners' short essays written in various TOP tests.
2.CSL students' writing assignments at MTC
3.CSL students' writing in the MTC achievement tests at each proficiency level

To facilitate corpus search, the learner corpus was further automatically tagged with a Chinese tagger called CKIP (Chinese Knowledge Information Processing) tagger developed by Academia Sinica, Taiwan. The POS-tagged CSL corpus is useful for research and teaching. In addition to the learner corpus, a web concordancer which has several different search options was also developed. This web concordance allows users retrieve specific words and phrases from CSL learner corpus. Thus, various CSL learners' errors can be retrieved and studied more easily and systematically. Furthermore, the POS-tagged learner corpus can be used to search for collocates used by CSL learners. Furthermore, because these data were produced by learners from various L1 language backgrounds, users of the web-based corpus system can also find errors and patterns produced by CSL learners from different native language backgrounds. The availability of this CSL learner corpus and the web concordancer should be able to help more researchers uncover CSL interlanguage patterns and conduct various types of research.

Building a Translation Learner Corpus: Methodology and Application

Wallace Chen, PhD
Graduate School of Translation, Interpretation, and Language Education
Monterey Institute of International Studies
Monterey, CA, U.S.A.

Keywords: computer-assisted translation (CAT), search engine, do-it-yourself corpus, concordance, reference corpus (RC), translation learner corpus (TLC), sight translation

Translation students in the current Internet era rely heavily on search engines such as Google(Shei 2010)or computer-assisted translation (CAT) tools (Aiken and Balan 2011) to find readily available answers for many translation problems they encounter. Although Google provides lightning-fast and ultra-rich search results, it does not necessarily present the most appropriate and relevant translation solutions to students. On the assessment side, translation students mostly look for guidance and suggestions from their instructors, paying less attention to peer performance that might shed some light on the improvement of translation skills. This study attempts to address these two issues by exploring a methodology to build do-it-yourself corpora and query them with concordancers. Specifically, the corpora referred to in this study are 1) a reference corpus (RC) of speeches and domain-specific documents in English, and 2) a translation learner corpus (TLC) consisting of English transcripts made by trainee translators in their Chinese-English sight translation classes. Detailed procedure of building these corpora as well as their applications will be presented and discussed. By consulting TLC, learners will be able to systematically identify error patterns, individual styles and issues, length of delivery, and the possible connections between lengthy delivery and language use. RC, on the other hand, serves as a benchmark against which translator trainers can compare learners' sight translation outputs at various linguistic levels, including collocation, terminology, idiomatic expression, verification of intuition, translation equivalent, target language patterns and new expressions. It is further suggested in this presentation that the corpus-assisted approach to sight translation evaluation offers an empirical tool to complement traditional approaches that are based on intuition, personal experience, subjective judgment and restricted knowledge of subject matters. By adopting a corpus-assisted approach to learning sight translation, students will be able to access more linguistic choices in addition to what they can find on Google or from other conventional resources.

References

Aiken, M. and S. Balan (2011). "An Analysis of Google Translate Accuracy." Retrieved July 12, 2011, 2011, from http://translationjournal.net/journal/56google.htm.

Shei, C. (2010). "On the Issue of Norms in Translating: the Function of Web as Corpus." Compilation and Translation Review **3**(2): 163-202.

Exploring the Limits of Phraseology

Winnie Cheng

Department of English
The Hong Kong Polytechnic University
Hong Kong SAR, P. R. China

Keywords: specialised corpora, concgram, meaning-shift unit, organisationally-oriented phraseologies, collocational frameworks

This paper adopts a very inclusive definition of phraseology to include all the patterns of associated words to be found in corpora (and individual texts). Such a definition of phraseology might require some substantiation in order to convince those who prefer a narrower definition. Through the presentation of examples of various kinds of phraseology, this paper sets out to describe the potential for recalibrating what constitutes phraseology, especially if greater attention is given to phraseological variation.

In this paper, two online specialised corpora are used, namely the 7.3-million-word Hong Kong Financial Services Corpus (HKFSC) and the 9.2-million-word Hong Kong Engineering Corpus (HKEC), hosted by the Research Centre for Professional Communication in English (RCPCE) and are freely available at http://rcpce.engl.polyu.edu.hk/. These profession-specific corpora were compiled with the help of a number of professional associations and individual professionals from the financial services and the engineering sectors in Hong Kong. For example, the Hong Kong Institution of Engineers, the Hong Kong Institute of Certified Public Accountants, the Hong Kong Securities Institute, government departments, and private organizations provided advice in terms of the kinds of texts to collect and the proportions to assign to each genre. One of the reasons for compiling the corpora was to inform professionals and learners of English for specific purposes (ESP) about these corpora and to train them to be able to interpret the results of their searches and use the results to inform their own writing.

The online corpora can be searched using tailor-made software, ConcGramOnline written by Chris Greaves, which has been designed to retrieve phraseology from the corpora, and other patterns of language use, which can then be interpreted.

This paper builds on previous studies of concgrams (see, e.g., Cheng, Greaves and Warren, 2006; Cheng, Greaves, Sinclair and Warren, 2009) which emphasised the importance of examining phraseological variation in addition to fixed phraseologies such as n-grams (i.e. contiguous sequences of two or more words such as *you know* and *I don't know*). Using the ConcGram software (Greaves, 2009), profession-specific phraseologies are investigated. The outputs of this software, concgrams (Cheng, Greaves and Warren, 2006), serve as useful raw data for finding phraseologies (i.e., two or more associated words), especially those phraseologies which have variation. The ability of the software to fully automatically retrieve variation in word

co-occurrences is important because it enables the researcher to look beyond n-grams when retrieving phraseologies from a corpus. The phraseological variation uncovered by the software includes both constituency variation when other words drop in between the associated words (e.g. ***rate increase*** and ***rate** of **increase***), and positional variation when the associated words are sequenced differently, but retain their canonical meaning (e.g. ***rate increase*** and ***increase** in the* rate).

The focus of this paper is on describing examples of phraseologies from the HKFSC and HKEC which are discipline-specific in that they are unique to, or more frequently found in, the financial services or engineering corpus. In addition to discipline-specific phraseologies, there are register-specific phraseologies which are specific to particular genres in the corpora; for example, annual reports and results announcements in the HKFSC and codes of practice and tender notices in the HKEC. Thus the findings have implications in terms of the ways in which such studies can also help to uncover the aboutness of professions, registers and genres, and even individual texts (Cheng, 2009).

The examples of phraseologies will draw on three major categories so far identified, namely lexically-rich phraseologies termed 'lexical items' or 'meaning-shift units' (Sinclair, 2004，2007), organisationally-oriented words and phrases, and collocational frameworks (Renouf and Sinclair, 1988) which frame lexical items. Each category illustrates that there are similarities and differences in the usage of these forms of phraseology when the specific disciplines and registers are investigated.

Acknowledgements

The research described in this workshop was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5459/08H, B-Q11N).

References

Cheng, W. 2009. *Income/interest/net*: Using internal criteria to determine the aboutness of a text. In Aijmer, A. (Ed.), *Corpora and Language Teaching. Am*sterdam/ Philadelphia: John Benjamins, 157-177.

Cheng, W., Greaves, C., Sinclair, J. McH, and Warren, M. 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, 30(2), 236-252.

Cheng, W., Greaves, C. and Warren, M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4), 411-433.

Greaves, C. 2009. *ConcGram 1.0: A Phraseological Search Engine.* Amsterdam: John Benjamins.

Sinclair, J. M. 2004. *Trust the Text*. London and New York: Routledge.

Sinclair, J. M. 2007. Defining the definiendom – new. (manuscript). Tuscan Word Centre, Italy.

Sinclair, J. M. and Renouf, A. 1988. A lexical syllabus for language teaching. In Carter, R. and McCarthy, M. (Eds.), *Vocabulary and Language Teaching*. London: Longman,140-158.

A Learner Corpus Study of Korean EFL Learners' Overgeneralization Errors

Incheol Choi

Department of English Education
Kyungpook National University, Korea

Keywords: learner corpus, causative alternation, overgeneralization

A language is not acquired by learning by memorization. Instead, learners acquire a rule system that underlies the language. What supports this idea most conspicuously is the overgeneralization errors made by language learners, which is exemplified in (1).

(1) a. I drinking my doll. [giving the doll a small bottle to drink] (Suppes, 1974; Nina 2;3;28)
b. I did fall my vitamin. [dropped the vitamin accidentally] (Lord, 1979)
(2) a. Maybe it's a building building up. [of a ruined castle in Vaison-la-Romaine] (Damon 2;9)
b. Where's your hand? Did it cut off? (Damon 2;10)

Overgeneralization has been considered a prevalent phenomenon not only in the first language acquisition, but also in the second language acquisition. Specifically, it was reported that learners with Chinese or Japanese L1 background overgeneralize the English causative alternation rule (Yip, 1990, 1995; Hirakawa, 1995; Ju, 2000).

To reveal whether Korean learners of English make the same type of overgeneralization errors, we carried out a learner corpus analysis. To do this, we used KELC (KNU English Learner Corpus) that contains essay writing tests consisting of 161,335 word and 838 texts by 396 children. Following Marcotte (2005), we extracted the sentences containing the following 10 intransitive and 10 transitive verbs from KELC.

(3) a. appear, stay, come, disappear, dive, fall, happen, jump, rise, run
b. catch, clean, cut, draw, drink, lose, pull, push, throw, wash

We also manually checked whether the sentences contained the errors. The total of the occurrences of the verbs was 1,992. Although we identified various types of errors, we decided that they have nothing to do with the overgeneralization of the causative alternation rule. Among them, the sentences in (4) involve the errors caused by the wrong omission of the prepositions.

(4) a. They arrived pizza restaurant. [KELC, 018madvt09st.txt]

b. Then we go to a my house next day! my friend come my house. [KELC, 304mpint09st.txt]

Another type of error that we considered is given in (5).

(5) a. Pizza were arrive in table and I ate a pizza. [KELC, 362madvt09st.txt]
b. Suddenly famous signer is came in the restaurant, may find him. [KELC, 395mintt09st.txt]

If the error sentences in (5) are considered the passivized versions of causative sentences, the sentences in (5) can be considered as examples of the overgeneralization error. However, the verbs in (5) do not have the passive verb forms. According to Yip (1990), the main cause of the overpassivization errors comes from the similarity of the underlying thematic structure between the unaccusative verbs and the passive verbs. However, our data showed that unaccusative verbs did not induce more overgeneralization errors than unergative verbs. Therefore, it is not likely that the be insertion errors are related to the overgeneralization errors in question.

After removing all the irrelevant errors, there remain only two sentences that can possibly be counted as overgeneralization errors:

(6) a. And then waiter is told about where they can stay the car. [KELC, 018madvt09st.txt]
b. We are some hot and many water drink and drink. [KELC, 049fpint03st.txt]

When this result is compared to the overgeneralization error rate of CHILDES investigated by Marcotte (2005), it seems obvious that Korean young learners make many fewer overgeneralization errors than native children (M = 0.0354, SD = 0.0937 for the CHILDES data compared to M = 0.0045, SD = 0.01862 for the KELC data).

A question that we need to address is whether the Korean learners actually acknowledge the alternation rule. This is because it is possible that the rarity of the overgeneralization errors can be ascribed to a lack of knowledge. To check this possibility, I carried out a grammaticality judgment test. The result showed that the young learners knew that the verbs in (7b) can appear both in transitive and intransitive sentences in contrast with the verbs in (7a) and (7c).

(7) a. appear, stay, come, jump
b. dry, melt, open, break
c. catch, clean, pull, lose

We suggest that the rarity of the overgeneralization errors was due to the Strict Lexical Conservatism in the sense of Baker (1979) and Fodor (1985). That is, although the Korean children acknowledged the causative alternation rule, they did not extend the rule to new forms because of their conservative learning attitude.

References

Baker, C. L. 1979. Syntactic Theory and The Projection Problem. *Linguistic Inquiry* 10, 533-581.

Fodor, J. D. 1985. Why Learn Lexical Rules? In *Proceedings of the Tenth Annual Boston University Conference on Language Development*, 25-27.

Hirakawa. M. 1995. L2 Acquisition of English Unaccusative Constructions. In *Proceedings of the 19th Annual Boston University Conference on Language Development* 1: 291-302.

Ju, M. K. 2000. Overpassivization Errors by Second Language Learners. *Studies in Second Language Acquisition* 22, 85-111.

Marcotte, J. 2005. Causative Alternation Errors in Child Language Acquisition. Doctoral Dissertation. Stanford University.

Yip, V. 1990. Interlanguage Ergative Constructions and Learnability. *CUHK Papers in Linguistics* 2. Hong Kong: Chinese University, 45-68.

Yip, V. 1995. *Interlanguage and learnability: From Chinese to English.* Amsterdam: Benjamins.

Modality in Philippine English: a DiachronicStudy

Peter Collins[1], Ariane Borlongan[2] and Xinyue Yao[3]

[1]*School of Languages and Linguistics, University of New South Wall*
[2]*Department of English and Applied Linguistics, De La Salle University*
[3]*School of Languages and Linguistics, University of New South Wales*

Keywords**:** modals, quasi-modals, Philippine English, diachronic corpus

Introduction

The diachronic study of Philippine English ('PhilE') has recently become possible through the compilation of a PhilE corpus ('Phil-Brown') at De La Salle University. The period of time defined by Phil-Brown (whose sampling period was late 1950s-early 1960s) and 'ICE-Phil', the Philippines component of the International Corpus of English (comprising texts sampled in the early 1990s), covers most of the period of time over which there has been general recognition of the existence of PhilE as a World English.

The study

Based on a selection of texts from Phil-Brown and ICE-Phil, we examined recent changes in the use of a set of modals (*may, might, must, ought, shall* and *should*) and quasi-modals (*be able to, be going to, be supposed to, have to, need to* and *want to*), investigating their frequency differences, genre variation, and semantic differentiation, and comparing the findings with those of Leech et al. (2009) for British and American English ('BrE' and 'AmE') of the same period.

Findings

Our findings generally suggest that PhilE does not pattern closely with either BrE or AmE, providing support for locating it in Phase 4 ('endonormative stabilisation') of Schneider's (2007) evolutionary scale. For example in broad frequency terms the six modals examined have, as a set, declined at a rate that is significantly slower than that in AmE and even slower than that in the typically conservative BrE. We also occasionally find patterns that are distinctive to PhilE at the micro-level of individual items, such as the considerably milder tendency towards monosemy with *may* and the high frequency of *shall* in PhilE. Other findings suggest that PhilE may not yet be ready to completely renounce its exonormative allegiance to its postcolonial 'parent', evidencing a higher rate of increase for the quasi-modals than AmE, which has enabled it to attain a comparable frequency to the latter. What this may suggest is that PhilE has been striving to 'catch up' with AmE over this period.

Some of our findings remain resistant to interpretation until we consider the behaviour of the modals and quasi-modals in specific genres. For example the relative mildness of the decline of the modals in PhilE might at first blush be taken to suggest that PhilE is more conservative than the two supervarieties. However upon closer inspection we find the apparent conservatism to be genre-specific rather than across-the-board, due primarily to the strong support for the modals in the Philippine press.

References

Leech, G., M. Hundt, C. Mair and N. Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Schneider, E. 2007. *Postcolonial English: Varieties of English around the World*. Cambridge: Cambridge University Press.

Recent Change in Non-Present Perfect Constructions in British and American English

Peter Collins[1] and Xinyue Yao[2]

[1]*School of Languages and Linguistics, University of New South Wales*
[2]*School of Languages and Linguistics, University of New South Wales*

Keywords**:** non-present perfects, grammatical change, British and American English

Introduction

A number of recent studies of grammatical categories in English have identified regional and diachronic variation in the use of the present perfect, suggesting that it has been losing ground to the preterit from the 18th century onwards (Elsness, 1997, 2009; Hundt and Smith, 2009; Yao and Collins, fc). This regional difference is perhaps most clearly manifested in the fact that British English (BrE) speakers tend to use the present perfect more frequently in contexts with 'stative' adverbs such as *already* and *yet* (e.g. *I have already finished*) than American English (AmE) speakers who would prefer the preterit (e.g. *I already finished*). Only a limited amount of research has been conducted on *non*-present perfects. More recently, Bowie and Aarts's (fc) study using the Diachronic Corpus of Present-Day Spoken English has found that certain non-present perfects had undergone a considerable decline in spoken BrE during the second half of the 20th century. However, comparison with AmE and across various genre categories has not been made.

The study

This study focuses on the development in the distribution of four types of non-present perfects (past, modal, *to*-infinitival and *ing*-participial) in BrE and AmE during the period from the early 1960s to the early 1990s. A part-of-speech tagged version of the four members of the Brown family of corpora – Brown, Frown, LOB and FLOB – was used as the major source of data. Comparison was made along three dimensions: diachronic (1960s vs 1990s), regional (BrE vs AmE) and stylistic (among general prose, press, learned and fictional writing).

Findings

This study found, *inter alia*, that contemporary BrE has a stronger preference than AmE for non-present perfects for the two time periods examined, and that non-present perfects – in particular past and modal perfects – have undergone a more significant decrease in AmE than in BrE. The non-present perfect thus takes its place alongside a number of grammatical categories responsible for widening the gap between the grammars of BrE and AmE. It is proposed that the decline of past perfects can be ascribed to a growing dispreference for past time references in various genres (particularly academic writing), which is related to long-term historical shifts associated with the underlying communicative functions of these genres. The decline of modal perfects, on the other hand, is more likely to reflect the influence of the general decline of modal auxiliaries in English.

References

Elsness, J. 1997. *The Perfect and the Preterit in Contemporary and Earlier English*. Berlin, New York: Mouton.

Elsness, J. 2009. The present perfect and the preterit. In G. Rohdenburg, G. and Schlüter, J. (Eds), *One Language, Two Grammars? Differences between British and American English*. Cambridge: Cambridge University Press, 228-45.

Hundt, M., and N. Smith. 2009. The present perfect in British and American English: Has there been any change, recently? *ICAME Journal*, 33, 45-63.

Yao, X., and P. Collins. fc. The Present Perfect in World Englishes. *World Englishes*.

Rules for POS Tagging of the Bengali Text Corpus

Niladri Sekhar Dash, Payel Dutta Chowdhury, and Amrita De

Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

Keywords: principles, rules, POS tagging, morphology, semantics, context, Bengali

1. Introduction

POS tagging of a natural text corpus is not an easy task. The thing that we need to keep in mind is that POS tagging, done manually or automatically, has to be carried out on a text corpus, which contains large number of words which are actually occurring within sentences with specific syntactic functions and semantic values assigned to them. Therefore, it is not easy to identify the specific POS of words, until and unless actual syntactic roles of words are properly understood and defined. Moreover, there are several issues related to POS tagging, such as, *text sanitation, text normalization, tokenization, orthographic error correction, spelling error correction, real word-error correction, grammatical error removal,* punctuation errors removal, etc. (Dash, 2009: 40-42). These issues lead us to propose a few rules, which we followed when we tagged words in the Bengali text corpus (Dash, 2011).

Rule 1: Tagging should be done at sentence level

POS tagging should be carried out at sentence level only – not on words standing alone in syntagmatic or paradigmatic distribution in different lexical databases. A particular POS tags to words should be done after evaluating their linguistic roles in sentences where they actually occur. This is mandatory because a word is often found to be used in different part-of-speech from the part-of-speech assigned to it in a dictionary (Dash, 2004).

Rule 2: Text should be normalized before

All broken word strings should be joined together before these are put to POS tagging (Sproat *et al.,* 2001). In many situations it is observed that a word is broken into two parts where the first part is a base and second part is a particle, case marker, or enclitic, e.g., *pere chila > perechila, kathā gulo > kathāgulo, meye der > meyeder,* etc.

Rule 3: Words should be tokenized before

It observed that sometimes some separate words are orthographically joined with their immediately preceding and succeeding words. These words have to be separated from their neighbouring members before these are used for POS tagging. For instance, *Rāmo Sītā > Rām o Sītā* "Ram and Sita"*, gelennā > gelen nā* "did not go", etc. Also, white space needs to be provided consistently before and after each word to ensure separate linguistic entity of a word.

Rule 4: Exact POS tag should be assigned

A word should invariably be tagged at the POS level exactly in what part-of-speech it is used in sentence. There should be no confusion or controversy in this regard. No attention should be given to lexicographic status or etymological antiquity of the words. For instance, in a Bengali sentence like *sonālī swapna dekhte bhālobāse* "Sonali loves to dream" the word *sonālī* should be tagged as a noun, and not as an adjective, even though *sonālī* "golden" is an adjective in standard Bengali dictionary.

Rule 5: Context should carry utmost importance

It is not at all advisable to POS tag words solely based on POS categories as proposed in dictionaries and grammars, as it may lead to problems in identification of actual POS roles of words in a text. Therefore, tagging of words should be entirely context-based and this will instruct and guide a POS annotator about how words are to be tagged in specific contexts taking into consideration lexical, semantic, and syntactic functions of words.

Rule 6: Existing POS categories should be used

It is advisable that POS tagset for a language should be designed in accordance with existing and accepted set of parts-of-speech, which has been understood for generations by language users. Additional POS category can be assigned only when it is found that accepted POS tagset is not adequate enough to address new functions of words noted in texts.

Rule 7: Support system should identify MWUs

There should be a support system for identification of multiword expressions in corpus. The examples of multiword units in Bengali include compound words, idiomatic expressions, complex verb forms, and proverbial expressions etc. These should be tagged as *chunks* as there is valuable lexicosemantic information involved in these forms, which asks for separate investigation vis-à-vis treatment.

Rule 8: Multi-tagging approach should be avoided

Multi-tagging approach should be strictly avoided. Although it may appear that a particular word can have more than one POS tag, one should invariably assign that particular tag, which the word under study exerts in particular context of its occurrence.

Rule 9: MP must be separated from POS Tagging

Morphological processing and POS tagging of words are two different processes and therefore should be treated separately. In fact, information extracted from morphological analysis may be used in POS tagging of word as well as in lexical form generation, machine learning, information extraction, and parsing. But it should never be mixed up with the task of POS tagging.

Rule 10: Hyphenated words needs special attention

We have to be careful in POS tagging of hyphenated words, as there are several morpho-semantic issues involved in them. In case of words where a formative element (e.g., inflection, particle, case marker, etc.) is separated from the word with a hyphen, it is better to

tag the entire hyphenated word as a single lexical unit, since these formative elements are actually the part of the base form, e.g., *ho-yāṭ* "what", *mā-i* "mother herself", *kālidās-er* "of Kalidas", etc.

Rule 11: Ambiguity should be resolved

There are words, which are ambiguous in sense denotation when used in text, e.g., *bhābe* can be used as a FV, NN, and Particle. So POS tagging rules should explicitly spell out the tagging conventions to be adopted for these ambiguous words.

Rule 12: Manual verification and validation

POS tagging should be done together by at least three experts who are well versed in morphology, grammar, morphosyntactic rules, and syntax of the language. This will provide the tagged corpus the much needed authenticity derived from 2:1 ratio of tag assignment for accuracy (Leech, 1993).

2. Conclusion: Value of POS Tagged Corpus

The importance of a POS tagged corpus in enormous in language description, natural language processing, and language technology (Leech, 1997). It indicates that we should initiate in this direction to develop POS tagged corpora for Bengali with two goals: design maximally accurate tagset to increase rate of accuracy of the POS tagged data, and develop the POS tagged corpora in a large scale covering all text types for future linguistic works.

References

Dash, N.S. 2004. Text annotation: a prologue to corpus processing. *Indian Journal of Linguistics*. 23(1), 71-82.

Dash, N.S. 2009. *Corpus-based Analysis of the Bengali Language*. Saarbrucken, Germany: VDM Publications.

Dash, N.S. 2011. Part-of-Speech (POS) tagging in Bengali written text corpus. To appear in *Journal of Linguistics and Technology (SNLTR Journal)*.

Leech, G. 1993. Corpus annotation schemes. *Literary and Linguistic Computing*. 8(4), 275-281.

Leech, G. 1997. Introducing corpus annotation. In, Garside, R., Leech, G. and McEnery, A. (Eds.) *Corpus annotation: Linguistic information from computer text corpora*. London: Longman, 1-18.

Sproat, R., Black, A.,Chen, S.,Kumar, S.,Ostendorf, M. and Richards, C. 2001. Normalization of non-standard words. *Computer Speech and Language*. 15(3), 287-333.

Lexical Generativity of Bengali Prefixes: A Corpus-based Investigation

Niladri Sekhar Dash and Baisakhi Topdar

Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

Keywords: Bengali, prefix, prefixation, lexical generativity, morphosemantics, meaning

1. Introduction

The present investigation aims at understanding the nature and patterns of use of prefixes in word formation in Bengali. It deals with a set of prefixes, a large list of Bengali words, and a set of word formation rules that operate at morphological level manipulating basic structures of words involved in prefixation to generate new words. It is a dynamic linguistic process in Bengali. As word-formation elements, they carry tremendous importance in understanding surface forms of words in Bengali, as understanding form and function of prefixes helps in systematic exploration of lexicosemantic roles of words in the context of formation of new words in the language. Also they help in interpretation of syntactic roles of words in transferring linguistic messages embedded within sentences.

2. The Phenomenon

In Bengali, the number of prefix is far less than the number of suffix. Prefixes are tagged at the initial position of words to form new words (Spencer, 1991:13). In case of word formation, although prefixes are quite productive in nature as these are able to be tagged with almost all nouns and adjectives, these are not recursive, i.e., these are not repeatedly used with a same word again and again. Examples collected from Bengali corpus shows that these are used before nouns, adjectives, verbal nouns, and participial forms of verbs in the following manners to produce new words in the language.

(a) prati-$_{[Prefix]}$ + din$_{[Noun]}$ “day” = pratidin “everyday”
(b) nim-$_{[Prefix]}$ + rāji$_{[Adjective]}$ “willing” =nimrāji “not willing”
(c) ni□-$_{[Prefix]}$+ gaman$_{[Verbal Noun]}$ ‘going”= nirgaman “excretion”
(d) bi-$_{[Prefix]}$ + gata$_{[Participial Verb]}$ “gone” = bigata “bygone”

3. Types and Forms of Bengali Prefixes

Prefixes are tagged immediately before a word to generate a new word. Due to this phenomenon, it is sensible to consider prefixes as integrated parts of words and treat prefixed words as single word units. The total number of prefix found to be used in Bengali word-formation is 73. Among these, prefix inherited from Sanskrit is 21, native prefix (i.e., prefixes acquired from native Bengali source or neighboring languages) is around 31, and prefix borrowed from foreign languages (i.e., Arabic, Persian, English and others) is around 21 (Tagore, 1910: 220, Thompson,

2010: 504).

4. Frequency of Use of Bengali Prefixes

The frequency of use of affixes with Bengali words shows that the percentage of use of Bengali words without affix is less (23.09%) than percentage of words with attached affix (76.91%). Moreover, among affixed Bengali words, percentage of use of words attached with suffix is higher (41.35%) than percentage of words attached with prefix (14.44%) and words tagged with prefix and suffix (21.12%). However, the usage patterns of prefix with Bengali words, if compared with that of suffixes, is more diverse and productive in the sense that their usage is invoked when there arises a need for generating words of different part-of-speech and meaning to address novel linguistic requirements. In a simple count, there are nearly 300,000 words in Bengali (Dash 2009) out of which nearly 65,000 words are affixed, and out of 65,000 words, nearly 15,000 words are attached with prefixes. In dictionary, the highest number of words are attached with prefix *a-* (2100), followed by *bi-* (1400), *sam-* (1000), *pra-* (940), and *su-* (920). In the corpus, the prefix *pra-* (15%) has the highest frequency of occurrence followed by that of *su-* (11%), *a-* (10%), *bi-* (9%),*sam-* (8%), *anu-* (7%), *uṯ-* (6%), *upa-* (5%) and *ā-* (4%). The frequency information of prefixes obtained from a dictionary does not always reflect on their frequency of usage in actual texts. A particular prefix may have limited number of entries in dictionary, but those limited entries can have high frequency of use in actual texts, e.g., *pra-* and *su-*.

5. Nature of semantic Change of Words with Addition of Prefix

From analysis of a list of prefixed words it is seen that although prefixes are usually class-preserving, they are invariably meaning-changing. Due to inherent semantic potentials, prefixes are able to denote new meanings of words. In fact, prefixes play crucial roles in formation of new words with new shades of meaning in Bengali. Generally, newly formed words denote a meaning which is either opposite to the meaning of old words or an enhancement of meaning of the existing words. For instance, *gata* "gone" changes its meaning when different prefixes are attached to it.

6. Change in POS of Words with Addition of Prefix

Usually addition of a prefix with word does not cause change in original part-of-speech of word. That means newly formed word remains to same part-of-speech to which original word belongs. However, since Bengali affixes are productive in nature, in some situations, addition of certain prefixes changes part-of-speech of words, as the following examples show.

nā-	+	bālak$_{[Noun]}$	"boy"	>	nābālak$_{[Adjective]}$	"immature"
sā-	+	bālak$_{[Noun]}$	"boy"	>	sābālak$_{[Adjective]}$	"adult male"
phi-	+	bachar$_{[Noun]}$	"year"	>	phibachar$_{[Adverb]}$	"every year"
har-	+	roj$_{[Noun]}$	"day"	>	harroj$_{[Adverb]}$	"every day"

The list shows that there is a feature of selection restriction in use of prefixes with words in generation of words of different part-of-speech, as only a limited few (not all) are allowed to take part in this kind of operation. Although this feature determines which prefixes can be attached with words to produce words of different part-of-speech, it is not known why and how these prefixes are selected for this particular kind of task. Perhaps, we need to explore into the form and meaning of the prefixes as well as of words to know why these prefixes entertain a special kind of

privilege to which other prefixes do not have any access. A detailed analysis of each of the prefixes needs to be done in order to understand this aspect of the language.

7. Addition of Multiple Prefixes in Formation of Words

In case of formation of new words in Bengali, prefixes are quite productive in nature, as they are able to tag with almost all nouns and adjectives. However, they are not recursive in nature in the sense that a particular prefix is not repeatedly used with the same word again and again to generate several new forms. Moreover, not more than one prefix is usually used with a particular word in a systematic order to generate another new word. There are only a few words where more than one prefix is tagged to produce a new word, as the following examples show.

an- + ā- + gata : anāgata "not came yet"

a- + pari- + bartita : aparibartita "unchanged"

8. Conclusion: Relevance of the Study

The present study has application relevance in descriptive linguistics, applied linguistics, and language technology. Learning to identify prefixes and knowing their forms and functions are of great value for expanding one's knowledge of grammar of a language. In-depth information about morpho-dynamic feature of prefixes supply knowledgebase to enhance linguistic creativity of language users as well as a good scope for expanding their vocabulary in the language. Since Bengali is a language that has many borrowed words in its vocabulary, a reference to etymology of commonly used prefixes will do good for the users in understanding their meanings and functions.

References:

Dash, N.S. 2009. *Corpus Based Analysis of the Bengali Language*. Saarbrucken, Germany: VDM Publication.

Spencer, A. 1991. Mo*rphological Theory*. Oxford: Basil Blackwell.

Tagore, R. 1910. Sh*abdatattva (Bengali Philology)*. Kolkata: Viswabharati Publications.

Thompson, H. 2010. *Bengali : A Comprehensive Grammar*. London and New York: Routledge (Taylor and Francis).

Co-constructed and Negotiated Apologies.

Contributions of Corpus Linguistics to the Study of Speech Acts

Gusztav Demeter, Ph.D.

Department of English
Case Western Reserve University, Cleveland, OH, USA

Keywords**:** apologies, speech acts, spoken discourse, co-construction

Apologies have been studied using different perspectives. Most instruments used in such studies, such as Discourse Completion Tests, role-plays, and written questionnaires, have neglected the fact that apologies occur in an interactional context, and that there are features of interactional discourse that may contribute to the construal of apologies (Deutschmann, 2003). A corpus linguistics methodology (McEnery and Wilson, 1996) would give a more accurate description of how apologies are produced in actual language use. The aim of the present study is to show that a corpus linguistics approach to apologies allows for identifying uses of apologies that cannot be arrived at by analyzing elicited data.

The Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000; Du Bois et al., 2003; Du Bois and Englebretson, 2004, 2005) and the spoken section of the Corpus of Contemporary American English (COCA) (Davies, 2008), totaling 81,806,485 words, were analyzed to establish the forms and functions of apologies occurring in these corpora. The most important finding is that apologies can be co-constructed by several participants in the interaction. This can be seen in the following example:

KENDR: A cookie baking set.
MARCI: Al right.
MARCI: Al right.
KENDR: Mm.
KEVIN: Rubber Maid.
MARCI: Oh.
MARCI: Let me see it.
KEN: &=laugh &=laugh &=laugh.
KEVIN: You can't squash it.
KENDR: Mm.
MARCI: Oh...
KENDR: Rubber Maid.
MARCI: neat.
KEVIN: Twelve pieces.
KENDR: Yay.
KEVIN: &=GASP.
KEN: That's XX...

MARCI: Oh that's X.
KENDR: Wow.
KEVIN: Oh that includes all the teaspoons though.
MARCI: In blue.
KENDR: In blue
that's not my color.
WENDY: It's not green.
I'm sorry.
KENDR: &=tsk.
MARCI: They don't come in green.
KEVIN: We bought it before you had an apartment.
KENDR: No my plates are blue
that's okay.
(SBCSAE, SBC013 Appease the Monster)

In this example, Kendra receives a cookie baking set as birthday present. However, the teaspoons are blue, while Kendra's favorite color is green. As Wendy is aware of the fact that Kendra's favorite color is green, she apologizes for the fact that the color of the teaspoons is not green, that is it does not meet Kendra's expectations, by using the apology "I'm sorry." However, the interaction did not stop here. Following the non-linguistic verbal response by Kendra, Marci believes that Wendy's apology was not enough, and steps in to complete the apology stating that the set does not come in green. Moreover, Kevin also contributes to the apology by providing an explanation, "We bought it before you had an apartment." The apology is therefore co-constructed, as more than one speaker participates in the speech act. Besides being co-constructed, this apology is also negotiated. As Kendra is not satisfied with Wendy's apology, the other participants elaborate it until Kendra is satisfied and utters "*that's okay.*"

In conclusion, the results show that corpus linguistics can contribute to a great extent to the study of speech acts. There are uses of apologies, such as the ones reported here, that can only be found through a usage based approach, as traditional data elicitation does not provide the necessary interaction in which speech acts usually occur.

References

Davies, M. 2008. The Corpus of Contemporary American English (COCA): 400+ Million Words, 1990-present. http://www.americancorpus.org

Deutschmann, M. 2003. *Apologising in British English*. Unmeå: Umeå Universiteit.

Du Bois, J. W., Chafe, W. L., Meyer, C. and Thompson, S. A. 2000. *Santa Barbara corpus of spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A. and Martey, N. 2003. *Santa Barbara corpus of spoken American English, Part 2*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W. and Englebretson, R. 2004. *Santa Barbara corpus of spoken American English, Part 3*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W. and Englebretson, R. 2005. *Santa Barbara corpus of spoken American English, Part 4*. Philadelphia: Linguistic Data Consortium.

McEnery, T. and Wilson, A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

What Linguistic Features Determine the CEFR Reading Levels?

Kosuke Doi

Department of Applied Linguistics
Tokyo University of Foreign Studies, Japan

Keywords**:** CEFR, reading, criterial features, vocabulary, grammar

Introduction

The common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) is spreading its influence outside Europe. The CEFR defines communicative proficiency at six levels - A1 and A2 (Basic User), B1 and B2 (Independent User), and C1 and C2 (Proficient User), according to illustrative descriptors given in functional terms.

It seems to be the case that the CEFR works as a beneficial tool for those who are involved in language education, but there are some problems. Little (2007) pointed out that the CEFR is not language-specific. It describes the communicative function that learners can perform at different proficiency levels, but does not specify how those functions might be realized in a particular language. For practical use of the CEFR, it is necessary to determine the linguistic features which are peculiar to each language, and relate them to the CEFR levels (Milanovic, 2009).

This study focuses on reading and aims to extract criterial features for CEFR reading levels of English. Based on Alderson et al. (2006), this study involved the following steps:

- identifying tests and tasks that have been incontrovertibly scaled on the CEFR;
- developing measures of the linguistic features of texts and tasks that previous research has shown to be relevant to defining difficulty, independently of the CEFR; and
- applying such measures experimentally to the texts and tasks identified in the first step to see to what extent analysis of the linguistic features of such texts and tasks can predict CEFR levels (p. 20).

Method

The texts used in this study were reading sections of the Cambridge ESOL Main Suite, which were administered between 2005 and 2010. The exams are aligned with the levels described by the CEFR (Cambridge ESOL, 2011). The Corpus of the exams consists of 49 exams and contains 130,124 words.

The corpus was quantitatively analyzed with the help of software. Various indices on lexical richness were calculated using RANGE (Heatley et al., 2002). The measures include the number of word types, type-token ratio (TTR), Guiraud Index (GI), Corrected TTR, advanced TTR, and advanced GI. To calculate syntactic complexity, L2 Syntactic Complexity analyzer (L2SCA; Lu,

2010) and Link Grammar Parser (Sleator and Temperley, 1991) were used. L2SCA computes the following 4 types of syntactic complexity indices: length of production unit, sentence complexity, subordination, coordination, and particular structures. Link Grammar Parser explains syntactic structures based on the relation between words. Kruskal-Wallis rank sum test and ratio test were employed to ascertain the differences of the measures among CEFR levels.

Results

The results indicated that vocabulary difficulty along with syntactic complexity characterized the CEFR reading levels. The proportion of word types covered by the list of the first 1,000 most frequent words generated from the British National Corpus demonstrated significant between-level differences except between A2 and B1. The other measures of vocabulary demonstrated significant differences only between a higher level and a lower level (e.g., C2 and A2).

As to syntactic complexity, the mean length of units tended to be longer at higher levels. This can be explained in part by the increase of complex nominals. Statistically significant differences were found between the A level and the B levels as well as between the B levels and the C levels for the proportions of pre-noun adjective and post-noun modifiers. Another reason for the increasing length is the use of subordinations. Although significant differences were found only between a higher level and a lower level, the number of subordinations per units increased as the level went up.

Acknowledgements

This study is based on my MA thesis. I would like to express my fullest gratitude to my supervisor, Professor Tono.

References

Alderson, J. C. et al. 2006. Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3 (1), 3-30.

Cambridge ESOL. 2011. Common European Framework of Reference (CEFR). Retrieved from http://www.cambridgeesol.org/about/standards/cefr.html.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Heatley et al. 2002. RANGE and FREQUENCY programs.

Little, D. 2007. The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal,* 91, 645-655.

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15, 474-496.

Milanovic, M. 2009. Cambridge ESOL and the CEFR. *Research Notes ,* 37, 2-5.

Sleator, D. D. K. and Temperley, D. 1991. Parsing English with a Link Grammar. Retrieved from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/link/pub/www/papers/ps/tr91-196.pdf.

Visualizing a Corpus of Spoken Classroom Discourse

Paul Doyle

English Language Specialist,
English Language Institute of Singapore,
Ministry of Education

Keywords**:** visualization, classroom interaction, discourse patterns, action research

Introduction

In this paper, I develop a case for a data-driven approach to developing English language teachers' critical reflections on classroom talk to improve their pedagogic practices. In doing this, I take up the challenge put forward by Hargreaves (2007) that "teaching is not at present a research-based profession" and advocate a specific type of evidence for improvement, that is, corpora of teachers' and students' talk. Previous research has looked at the intersection of corpus linguistics with initial teacher education (O'Keefe and Farr, 2003) and feedback for teaching practice (Farr, 2010), but as yet, little if any research has been done on the use of corpus data *from* teaching to directly inform the teachers' own practice.

Background

Corpus data provide English language teachers with ways of viewing multiple examples of lexical and syntactical patterns, enabling them to explore phraseology, collocation and idiomatic language use in the context of their own and others' spoken classroom discourse. In addition, many corpora are now available online, and provide increasingly sophisticated search tools, thus widening the accessibility of corpus data to non-corpus linguists. Nevertheless, corpus data still presents challenges if it is to be interpreted and utilized in productive ways by practitioners rather than researchers.

In the paper, I will discuss the use of the Singapore Corpus of Research in Education (SCoRE) (Doyle and Hong, 2009) for enhancing teachers' awareness of their language use and teaching strategies such as questioning techniques through the presentation and discussion of patterns derived from classroom interaction data. SCoRE is a 2.3 million word corpus of spoken data comprising annotated transcripts of English, Science, Mathematics and Social Studies lessons from the fifth year of primary and third year of secondary schooling.

The classroom data was originally collected as part of a larger project designed to investigate the classroom practices and language in Singapore schools. As a result, methods were developed for exploring the classroom discourse of teachers and for annotating patterns in the discourse, such as the well-known Interaction, Response, Feedback or Evaluation (IRF / IRE) pattern of teacher student interactions (Sinclair and Coulthard, 1975; Cazden, 2001). The annotation scheme used was informed by the work of Cazden (2001) and adapted for the Singapore classroom context, where the presence of Singapore English alongside standard English forms requires more

careful analysis.

Data visualization

To make use of corpus data effectively for the purpose of improving teachers' practices, I have argued (Doyle, 2011) that more interactive and pedagogically-oriented interfaces to the rich language data found in corpora are required before these resources can be fully exploited by teachers. This requires the development of methods of data visualization (Spence, 2010; Mazza, 2009) that go beyond existing ways of viewing data, such as KWIC concordances or collocation grids, to create more sophisticated mappings of discourse patterns to pedagogic processes. Recent work in the field of data visualization (Yau, 2011; Freiberg, 2009; Few, 2009) provides a number of techniques, but how successful will these be in supporting researcher-practitioner dialogue?

Classroom research

The paper presents a case study of a collaborative project with a Singapore secondary English language teacher which evaluated the use of data visualization to prompt the teacher's critical reflection on and pedagogic interpretation of patterns in her classroom interactions with students.

I discuss how a visualization of teacher questions and follow up moves in annotated English language lesson transcripts was developed, and how the visualization was integrated with video recordings of the lessons so that the teacher could learn to 'see' the pedagogy patterns in the post lesson critical reflection interview.

In addition, I comment on the potential of combining this micro-level focus on one transcript with the macro-level focus on the corpus (SCoRE) as a means of refining teacher's critique of their own practice.

References

Cazden, C. 2001. *Classroom discourse. The discourse of learning*. Portsmouth, NJ: Heinemann.

Doyle, P. 2011. *Developing data visualization techniques for enhancing professional development: Viewing patterns of discourse and disciplinarity in classroom corpus data*. Final Report for Funded Research Project, OER38/08PD. Singapore: Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University.

Doyle, P. and Hong, H. 2009. *Compiling a multimodal corpus of educational discourse in Singapore schools*. Final Report for Funded Research Projects, CRP 20/04 JD, CRP 7/05 AL and CRP 13/05 AL. Singapore: Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University.

Doyle, P. 2011. Viewing Language Data Patterns: Data Visualization for Data-driven Learning. In Ho, C. M. L., Anderson K. T.and Leong, A. P. (Eds), *Transforming Literacies and Language. Multimodality and Literacy in the New Media Age*. London: Continuum, 149-166.

Farr, F. 2010. *The discourse of teaching practice feedback: A corpus-based investigation of spoken and written modes*. (Routledge advances in corpus linguistics). London: Routledge.

Feinberg, J. 2010. Wordle. In Steele, J. and Ilinksy, N. (Eds), *Beautiful Visualization. Looking at Data Through the Eyes of Experts*. Sebastapol, CA: O'Reilly Media, 37-58.

Few, Stephen. 2009. *Now you see it. Simple Visualization Techniques for Quantitative Analysis*. Oakland, CA: Analytics Press.

Hargreaves, D. 2007. Teaching as a research-based profession: possibilities and prospects. In

Hammersley, M. (Ed.), *Educational research and evidence-based practice*. London: The Open University and Sage, 3-17.

O'Keefe, A. and Farr, F. 2003. Using language corpora in initial teacher education: Pedagogic issues and practical applications. TESOL Quarterly, 37 (3); 389-418

Mazza, R. 2009. *Introduction to Information Visualization*. London: Springer-Verlag.

Sinclair, J. McH. and Coulthard, M. 1975. *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford: Oxford University Press.

Spence, R. 2009. *Information Visualization. Design for Interaction*. (Second Edition). London: Prentice Hall, Pearson Education.

Yau, N. 2011. *Visualize This. The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.

Language Statistics-Based Quality Assurance for Large Corpora

Thomas Eckart, Uwe Quasthoff and Dirk Goldhahn

Natural Language Processing Group
University Leipzig, Germany

Keywords**:** web corpora, large corpora, corpus evaluation, quality assurance

The WWW as a Resource for Large Corpora

With the growth of online resources the WWW became one of the most important sources for corpus building. The availability of massive amounts of text makes quality assurance a major challenge as corpora of up to millions of sentences can hardly be evaluated by hand. In addition sufficient language competence is not always available when dealing with a wide range of different languages. Especially crawling generic web pages for "smaller" languages often leads to inconsistent and heterogeneous corpora, including problems with encoding, inclusion of specific markup or false language elements. As a consequence statistical measures can be evaluated for their use as indicators for accurate corpora and as a first step to computer-assisted quality management.

Quality Management of Corpora

In many areas (including software engineering or project management) quality management is systematically used to improve the quality of results. Relevant stages of a long-time quality management include (among others) quality assurance and quality improvement (Rose, 2005). Therefore it was the goal to identify features that could be used as indicators for corpus quality as a basis for quality improvement procedures (including additional post-processing steps like correction or deletion of ill-formed data or adaption of used segmentation rules).

These features should address general properties of natural language texts and hence, be as language-and domain-independent as possible. Furthermore these features have to be measured automatically in an acceptable period of time.

There might be different reasons for measuring unexpected values:

- Poor preprocessing of the corpus: Leaving behind non-textual strings like markup, text in a wrong language or wrong character set
- Unexpected corpus content: Many repeated structures as in machine generated text, texts with one dominant subject area like religious texts
- Poor language quality: Often found in forums or in machine translated texts

Quality Measurement: Features

Besides looking on specific outliers (c.f. Eckart and Quasthoff, 2010) there is a variety of features that can be systematically used. These include:

- character statistics, especially for rare characters which might be a relict of failed character conversion
- typical length distribution for word, sentence, or paragraph length
- typical character or n-gram distributions
- untypical high frequency words
- conformity to well-known empirical language laws, like Zipf's law (Zipf,1972)
- typical correlations like correlation of word frequency with number of significant word co-occurrences (Dunning, 1993) or correlation of number of sentences and total number of significant word co-occurrences
- relative number of segments with untypical properties, like sentences without any stop words or with an extensive number of special characters

Examples

The following examples are taken from the Leipzig Corpora Collection (see *http://corpora.informatik .uni-leipzig.de/*). For this project corpora in more than 60 languages and in different writing systems were processed in a unified way (Eckart and Quasthoff, 2010) and were made available for search or download.

As an example, Figure 1 shows the sentence length distributions measured in characters for three different corpora. These include two accurate corpora (based on Hindi and Marathi newspapers) and a problematic corpus based on the Sundanese Wikipedia. The latter shows an apparent deformation compared to the typical distribution that can be found in corpora for most languages. For the Sundanese Wikipedia there are different reasons for the two peaks of the distribution: The first peak is due to a large set of sentences containing exactly 44 characters like

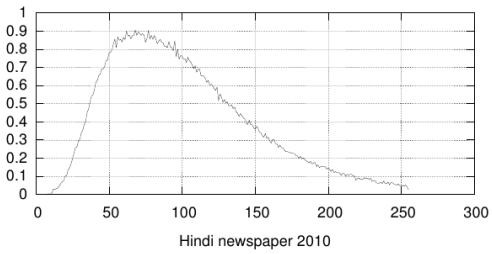
- Taun ka-1987 Maséhi dina Kaléndér Grégorian.
- Taun ka-1988 Maséhi dina Kaléndér Grégorian.

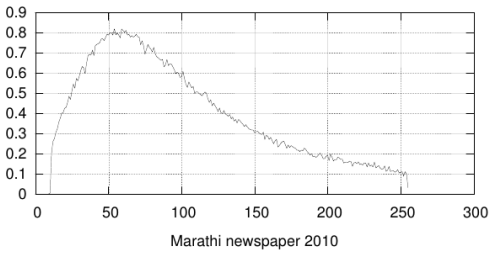
These sentences are nearly identical and could be removed before any statistical measurements. The second peak is due to the following kind of sentences:

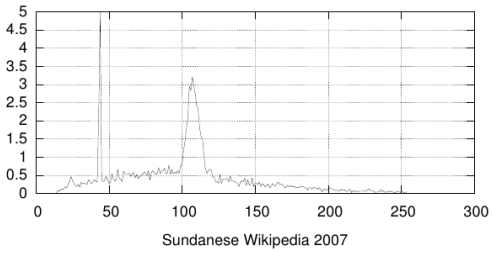
- Ancol nyaéta salasahiji désa di kacamatan Cinéam, Kabupatén Tasikmalaya, Propinsi Jawa Barat, Indonésia.
- Babakan nyaéta salasahiji désa di kacamatan Wanayasa, Kabupatén Purwakarta, Propinsi Jawa Barat, Indonésia.

These sentences are possibly machine generated, but they are more difficult to be detected using string similarity methods.

Figure 1: Sentence length distribution for three corpora (percentage for number of characters)







References

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1).

Eckart, T., and Quasthoff, U. 2010. Statistical corpus and language comparison. Workshop on *Building and Using Comparable Corpora*, LREC 2010, Malta, 2010.

Rose, K. H. 2005. *Project Quality Management: Why, What and How*. Fort Lauderdale, Florida: J. Ross Publishing.

Zipf, G. K. 1972. *Human behavior and the principle of least effort : an introduction to human ecology*. Hafner reprint, New York, 1972, 1st edn. Addison-Wesley, Cambridge, MA, 1949.

Ideology in News Reporting: A Multimodal Corpus-based Study

Roberta Facchinetti

Department of Foreign Languages and Literatures
University of Verona, Italy

Keywords**:** multimodality, visuals, news reports, print journalism, ideology

Scope and aim of the study

The present study has been prompted by recent research on (a) the expression of the writer's angle in news reports and (b) the interrelations between images and verbiage underlying the overall impact of the news piece.

Specifically, on the one hand, scholars have recently challenged the traditional pillars of 'impartiality' and 'objectivity' widely taken for granted as intrinsic of news reporting, and have moved towards considerations on 'perspective' and 'subjectivity', often rooted in the socio-geographical backgrounds against which the news discourse is produced. Indeed, although reporters have at their disposal a wide range of possibilities to express their views covertly, Adriaansen, van Praag *et al.* claim that "the way the media report the news has shifted from a mostly descriptive manner to an interpretative style". Fowler even claims that "news is (…) not a value-free reflection of 'facts'", a remark that dovetails with Pounds' statement that "all journalism is ultimately opinion journalism, in that it is always possible to detect signs of authorial stance even in so-called 'hard-news reporting' which is clearly marked as such" .

On the other hand, scholars have also started to investigate how far technological developments have impinged on 'the new face' of news-writing; thus, they have abandoned the notion of the news report as the exclusively linguistic output of the news-making process and have focused on co-textual aspects as well, particularly the interrelations between image and verbiage in both print and online newspapers. For example, Wells delves into the rhetoric of images, which contributes to shaping the narrative about the cause and consequences of events. In turn, Perlmutter and Smith Dahmenlink up images to their verbal captions by examining the websites, books and videos dedicated to exploring the 'moon-hoax' phenomenon, that is, the claim that the photos of 1969 Apollo moon landing were in reality a hoax. Finally, Caple and Bednarekstudy Image Nuclear News Stories (INNS), news reports whereby "a press photograph is combined with only a heading and a brief caption rather than an extended news report".

Starting from the above considerations, the aim of the present paper is to carry out a case study on the front pages of *The New York Times*, bringing to the fore the role played by the interrelationship between image and verbiage in conveying the writers'/subeditors' attitude to the news stories and, possibly, the underlying ideological background of the newspaper itself. The focus will be on the coverage by *The New York Times* of the birth and development of the European Union.

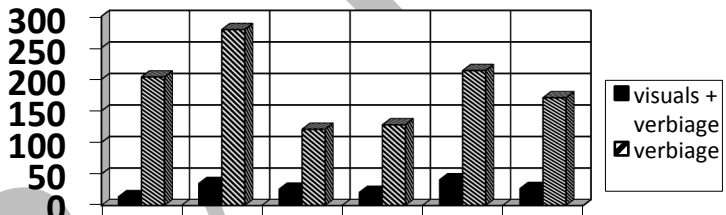
Corpus and methodology

To carry out the study, the PDF versions of the front pages of all the issues of *The New York Times* published between 1950 and 2009 were screened for the keyword 'euro', including, for example, 'Euro', 'Europe', 'European', and 'Eurodollar'; a second manual screening was carried out on such data to select only news items dealing with Europe integration/unity.

For each story, all news items were taken into consideration, including headlines, sub-headlines, visuals, captions, and body text. Indeed, the analysis of the corpus focuses particularly on the interplay between the 'syntactic implicitness' (Messaris and Abraham, 2001: 220) of the cover pictures, the related lexical context (captions and headlines), the associated body texts and finally the positioning of the whole piece in the front page layout.

Results

Figure 1 illustrates the quantitative results deriving from the analysis. The data show a high frequency of pure textual elements throughout all the decades analysed and only a limited number of news reports providing both verbiage and visuals:



	1950-1959	1960-1969	1970-1979	1980-1989	1990-1999	2000-2009
visuals + verbiage	14	35	26	21	41	27
verbiage	204	279	121	128	214	171

Figure 1: Raw data of the visuals and writings focusing on the EU-making process in *NYT* front pages.

The qualitative analysis of these data testifies to the fact that visuals and writing are distinct but equally significant and interrelated elements of the news-making process; furthermore, their interrelation strongly adds to the factual and evaluative information of the news reports and also, in certain cases, helps framing and articulating ideological messages. Specifically, the analysis has highlighted the following:

1. no regular direct alignment between visuals and writing: visuals do not always mirror what is reported in writing; this is frequently the case when the report highlights problematic aspects of the EU making process;
2. gradual, steady change in presenting the news on European integration by the *The New York Times* throughout the decades under scrutiny, particularly with reference to (a) the positioning of the pieces in the front page, (b) the focus on the role played by the United States in the EU-making process, and (c) the visuals selected, where only a set of countries are regularly given prominence.

The present analysis shows that newspaper discourse can no longer be viewed exclusively or mostly as a monolithic verbal text; on the contrary, it is the multi-faceted polyhedron whereby image, image-caption, headline, column, lay-out, and positioning in the page simultaneously contribute to the meaning-making process of the piece in a compositional way. Indeed, by now, the 'news piece' has turned into a 'news package' that calls for a holistic interpretation in order to be fully grasped.

References

Adriaansen, M. L., van Praag, P. and de Vreese, C. H.. 2010. Substance matters: How news content can reduce political cynicism. *International Journal of Public Opinion Research*, 22 (4), 433-457.

Caple, Helen and Monika Bednarek. 2010. Double-take: Unpacking the play in the image-nuclear news story. *Visual Communication*, 9 (2). 211-229.

Fowler, R. 1991. *Language in the News*. Abingdon: Routledge.

Martin, J. R. and White, P. R. R. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

Perlmutter, D. D. and Dahmen, N. S. 2008. (In)visible evidence: Pictorially enhanced disbelief in the Apollo moon landings. *Visual Communication*, 7 (2), 229-251.

Pounds, G. 2010. Attitude and subjectivity in Italian and British hard-news reporting: The construction of a culture-specific 'reporter' voice. *Discourse Studies*, 12 (1), 106-137.

Wells, K. 2007. Narratives of liberation and narratives of innocent suffering: The rhetorical uses of images of Iraqi children in the British press. *Visual Communication*, 6 (1), 55-71.

"110" in China: A Cognitive Approach

Xitao Fu

School of Linguistics and Applied Language Studies
Victoria University of Wellington

Keywords**:** "110", semantic value, syntactic value, metonymy, metaphor, Internet-ZH, *People's Daily*

Introduction

Emergency telephone numbers are adopted all over the world; they are typically a three-digit number, such as 111, 999, 911, and 112. In China 110, 119, and 120 are employed as the emergency numbers for police, fire and medical respectively.

However, recent years has witnessed a vast growing extended use of "110" from the emergency number for police to other areas such as governmental services and non-governmental services, i.e. those service numbers consist of the digits "110", generally ending with them, e.g., Food Safe 110; and "110" has been even further employed only as a label, i.e. those services do not consists of the three digits but they call themselves "×× 110", e.g.,Majhong 110.

Why can only 110 rather than 119 or 120 have such extended use in China? Why has the emergency phone number for police undergone such extended use only in China? What cognitive processes underlie this extension of 110 from the police to other domains? This paper is one of the attempts to answer these questions. It aims to uncover the underlying cognitive processes through a corpus analysis from the cognitive perspective.

Methodology and analysis results

In order to know exactly how "110" is currently used in China and what cognitive process(es) underlies the extended use, the search was specifically conducted in internet-based corpora from search engines such as the corpus Internet-ZH in Sketch Engine[5]and the People Website in the consideration of the advantage internet-based corpora bear, i.e. they are better than balanced/representative corpus such as BNC in terms of authorship, mode, audience, aim, and domain (Sharoff, 2006, 2007).,

The concordance of "110" as the query is 4674, but the hits for "110" as a numeral involving the emergency service total only 1344, which forms the corpus for my analysis. Then these hits were sorted out according to the syntactic value of "110" as used in the corpus. My findings show that "110" functions syntactically as noun, adjective, and adverb, with overwhelming quantity of

[5] See http://ca.sketchengine.co.uk/auth/preloaded_corpus/i-zh/ske/first_form, retrieved on 15 Nov, 2009. Internet-ZH is a Chinese web corpus with a size of 90 million words collected by Serge Sharoff. Such internet corpora are generally established in four steps: word selection, query generation, downloading and post-processing. Internet-ZH, as one of his internet-based corpora, was established in this way in 2005. It was tokenised and part-of-speech tagged using tools from North Eastern University, China. The following examples are taken from this corpus with kind permission. It is also available on his site at Leeds University, UK.

1017 hits as noun, 324 as adjective and 3 as adverb. The semantic values of 110 as a noun and adjective are shown below in Fig. 1 and Fig. 2.

In terms of semantic value, however, "110" as an adverb modifying "reporting" verbs (thus still in the aspect of quality) only involves the representation of the emergency phone number, indicating reporting to the police "by dialing '110'".

This analysis suggests that this extended use of "110", i.e. usage related to "110" as the emergency police number, is underpinned by the cognitive process of metonymy with EMERGENCY as the meaning prototype within the police domain. And this finding is different from Deignan (2006), which shows that metaphor dominates the conversion between different parts of speech of a word.

Besides this metonymy-motivated use, "110" is also found to be motivated by metaphor in its extension to other domains rather than police. Internet-ZH shows some examples. But a search in the People Website [6] found much more such extended use, involving almost every aspect of life, such as Environment, Charity, Agriculture, Tourism, Education, Science, Banking, Medical, Examination, psychology, etc. The analysis of the metaphoric extension of "110" indicates that the meaning prototype moves in accordance with context from EMERGENCY to INFORMATION, and to ASSISTANCE. But all these extended use is generally overseen by and developed from the global meaning of "110" which was formed when it was brought from a number domain to a special use.

Conclusion

Through the corpus analysis, this research shows the metonymy and metaphor underlie the extended use of 110 in China, with metonymy as the dominant cognitive process.

Acknowledgements: I am grateful to Serge Sharoff for sending his papers which were not otherwise available to me.

Fig. 1: The semantic representation chart of "110" as a noun

[6] http://search.people.com.cn/rmw/GB/rmwsearch/gj_search_pd.jsp, retrieved on 20 Nov, 2009. This website is an official governmental news site of China run by the People's Daily and People's Daily Online since 1997. The corresponding English website is http://english.peopledaily.com.cn/

Fig. 2: The semantic representation chart of “110” as an adjective

References

Deignan, A. 2006. The Grammar of Linguistic Metaphors. In Stefanowitsch, A. and Gries, S. T.(Eds.), Corpus-based Approaches to Metaphor and Metonymy. Berlin: Mouton de Gruyter, 106-122.

Sharoff, S. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In Baroni, M. and Bernardini, S.(Eds.), WaCky! Working Papers on the Web as Corpus. Bologna: GEDIT, 63-98.

Sharoff, S. 2007. Classifying Web Corpora into Domain and Genre Using Automatic Feature Identification. Cahiers du Cental, 5, 1-10.

Evaluation in Spoken Discourse:

An Exploratory Investigation of Stance in the MICASE

Delian Gaskell

Language Centre
Hong Kong University of Science and Technology, Hong Kong

Keywords**:** shell nouns, evaluative nouns, stance, discourse analysis, speech, corpus linguistics

Introduction

Specific expressions are used to portray the evaluative meaning behind the message we are trying to communicate. These expressions indicate what we think of what we are writing or talking about (Thompson and Hunston, 2000). For instance, some nouns carry an evaluative meaning that also communicates a specific stance that the writer or speaker has assigned to the text or discourse. Evaluative language is particularly prominent as a function of academic discourse, which is full of evaluation of other people's contribution to whatever field being assessed (R□mer, 2008). Stance has been the subject of extensive investigation within the context of academic writing (Barton, 1993; Biber and Finegan, 1988, 1989; Hyland, 1999, among others).

This exploratory study uses the academic spoken discourse represented in the Michigan Corpus of Academic Spoken English (MICASE) to examine how certain nouns, specifically those labelled by Hunston and Francis (1999) and Schmid (2000) as shell nouns, are used as epistemic stance markers. The aim of this study is to investigate the specific shell nouns that are used by speakers to mark epistemic stance. The research questions guiding this study are based on three themes: frequency, related patterns, and the anaphoric or cataphoric reference associated with epistemic stance markers.

Specifically, the research questions addressed in this study are:

1. Which are the frequent epistemic stance markers in academic speech found in the MICASE?

2. Is the frequency of *argument* the same or different in the two academic disciplines that make up the bulk of the MICASE?

3. Which word and grammar patterns are frequently associated with *argument* in academic speech?

4. How do these patterns appear to assist speakers to construct this type of stance?

5. a. Is *argument* used with anaphoric and cataphoric reference?

b. Are there phraseologies associated with this noun that are associated with anaphoric and cataphoric reference, and if so, what are the patterns of use when *argument* is used anaphorically and cataphorically?

The following text provides a description of how epistemic stance markers were identified from the large pool of potential evaluative nouns in order to answer question 1, then uses the noun *argument* as a representative example to shed light on questions 2-5.

Results

In order to answer question 1, an analysis of concordance examples retrieved from the MICASE was conducted to identify which of Hunston and Francis' 159 shell nouns (1999) were associated with a that-clause and related proposition. This reduced the number of potential evaluative nouns for investigation in this study to eight, which were, in order of descending frequency: *argument*, *feeling*, *sense*, *hypothesis*, *thought*, *concept*, *expectation* and *impression*. Question 2 focused specifically on the most frequent of these evaluative nouns, *argument*, to discover that it was used in spoken discourse as an epistemic stance marker more often within the broad discipline of Social Science (16.49 times per 100,000 words) in comparison to Physical Science (.73 times per 100,000 words). The next step in this analysis was to look at the words and grammatical features acting as modifiers of *argument*, specifically those that frequently occurred directly to the left of the noun when it was being used by speakers to provide an epistemic stance to an associated proposition. The most frequently occurring patterns were adjectives, articles and possessives, followed by the demonstrative determiners *that* and *this*. Question 4 focused on how the words and grammar patterns found directly to the left and right of *argument* were used in longer patterns to assist speakers to construct epistemic stance. The most frequently occurring longer patterns to the left of *argument* were found to be in association with the copular verb *BE* and the activity verb *MAKE*. The grammatical feature most frequently occurring directly before *BE* was pronoun usage, with either the demonstrative pronoun *that* (*that's*) or *thi*s (*this is*). Interestingly, in the majority of these instances, the speaker was using anaphoric reference with both the pronoun and *argument*, so the noun phrases being replaced by these demonstrative pronouns were also the propositions for which the speaker used *argument* to indicate the epistemic stance. Speakers in the MICASE produced frequent collocational strings with *MAKE* and *argument* in combination with the use of modals before the verb to create a sense of possibility or potential in their construction of epistemic stance. In terms of patterns of usage to the right, *BE* and the preposition *about* were used in conjunction with *argument* in cataphoric reference linking directly to an associated proposition. Question 5 required a manual review of the concordance examples associated with *argument* when it was used as an epistemic stance marker, specifically looking for the cohesive links speakers made before and after *argument* to the associated proposition. This revealed the answer to question 5a, that the distribution of reference was relatively balanced: 53% were cataphoric in comparison to 47% with anaphoric reference. In terms of question 5b, looking for the possibility of whether anaphoric and cataphoric reference are differentially associated with particular phraseologies, it was found that the seven categories of words that frequently occur directly before *argument* (as discussed in question 3) appear with both anaphoric and cataphoric reference with relatively equal frequency. However, there were two exceptions to this trend: the majority of the usage associated with the definite article was

cataphoric, while *argument* when paired with the demonstrative determiner *that* was used exclusively anaphorically.

References

Barton, E.L. 2003. Evidentials, argumentation, and epistemological stance. *College English*, 55 (7),745-769.

Biber, D. and Finegan, E. 1988. Adverbial stance types in English. *Discourse Processes*, 11, 1-34.

Biber, D. and Finegan, E. 1989. Styles of stance in English: lexical and grammatical marking of evidentiality and affect. *Text*, 9 **(**1), 93-124.

Hunston, S. and Francis, G. 1999. *Pattern Grammar*. Amsterdam: John Benjamins.

Hyland, K. 1999. Disciplinary discourses: writer stance in research articles. In Candlin, C.N and Hyland, K. (Eds), *Writing: Texts, Processes and Practices*. London: Longman, 99–121.

Römer, U. 2008. Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing. *Functions of Language,* 15 (1), 115-130.

Schmid, H.J. 2000. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton de Gruyter.

Thompson, G. and Hunston, S. 2000. Evaluation: an introduction. In Hunston, S. and Thompson,G. (Eds), *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1-27.

Are Interpreters Missing the Plot?

Developing a Corpus of 'Realia' Television Programmes to Test Interpreters' Knowledge of Idiomatic Language

Lynn Grant and Ineke Crezee

School of Language and Culture,
AUT University, Auckland

Keywords**:** realia television, corpus, idiomatic language, interpreter training

Introduction

The role of idiomatic language (single word or multiword items whose meaning cannot be accurately determined by adding together the literal meanings of the words involved) is sometimes overlooked in interpreter pedagogies (Baker, 2010). This is an issue that corpora can address (Aixelá, 1996; Bernadini, 2004).

This study reports on a corpus-based (Biber, 2009) investigation carried out by testing the awareness and knowledge of idiomatic language among a group of students studying interpreting. A small specialised corpus of approximately 120,000 words was developed comprised of transcribed 'realia' television programmes, that is programmes that might include some scripting but also involved people using natural everyday language in the course of their jobs. The aim was to select television programmes involving occupations where interpreters might be used such as police, customs, and medical situations.

This small specialised corpus included New Zealand and Australian programmes where these paramedics, customs or police officers were involved in conversations with members of the public. Television programmes such as *Border Patrol, Road Cops, Rapid Response, Borderline* were used, and episodes of these programmes were first videoed and then transcribed. The transcripts were searched for idiomatic language, which was then studied to find its function in the discourse. Following this, tests of idiomatic language were made based on selected individual programmes.

In the initial pilot project, students were presented with a dialogue and given multiple choice answers to match the most accurate meaning to the selected idiomatic language. Following the pilot project, a series of dialogues was transcribed and a progressively more difficult one was presented to interpreting students in Weeks 2, 6 and 10 of the twelve week course. This time the tests also included video clips showing excerpts of the language being used. For these tests, rather than selecting the best multi-choice answer, students were asked to explain the chosen examples of idiomatic language in their own words, aided by the audio and audiovisual technology..

Results

The study showed that students were often not aware of the idiomatic meanings of some

commonly used language in real life settings. But when shown a video clip of the language being used in the television excerpt, the audiovisual information helped some of them to deduce some of idiomatic meanings in context.

Initially some students claimed 'never to have heard of' the idiomatic language they were being tested on. After raising their awareness of it, however, the same students reported that they now 'heard it all the time'. The study therefore showed that it is important for trainee interpreters to be aware of this language as unfamiliarity with it may mean them 'missing the plot' in the spoken language they are interpreting.

Results of the study have led to recommendations for interpreter education pedagogies. This includes basing interpreting practice dialogues on the television transcripts to see if students could:

a)identify the idiomatic language themselves and accurately explain it, plus

b)develop a culture specific approach to their interpreting.

Further research is planned to see if the training of interpreters can be improved by increasing their recognition, knowledge and understanding of this type of what is often culture-specific idiomatic language.

References

Aixelá, J. F. 1996. Culture-specific items in translation. InÁlvarez, R.and Vidal, M. C. (Eds.),*Translation, power, subversion*. Clevedon, U.S.A.: Multilingual Matters, 52-78.

Baker, M. 2010. *In other words: A course book on Translation*. London, England: Routledge.

Bernadini, S. 2004. Corpus-aided language pedagogy for translator education. In: Malmkjaer, K. (Ed.) *Translation in Undergraduate Degree Programmes.* Amsterdam, the Netherlands: John Benjamins, 17-30.

Biber, D. 2009. A corpus-driven approach to formulaic language in English, Multiword patterns in speech and writing. *International Journal of Corpus Linguistics,* 14(3), 275-311.

The Revision of the Collocations in Some Entries of タイ日大辞典 with ThaiWaC

Keisuke Hashimoto

Department of Foreign languages,

Tokyo University of Foreign Studies, Tokyo, Japan

Keywords: ThaiWaC, lexicography, Thai, collocations

Introduction

Nowadays, it is widely acknowledged that corpus analysis is advantageous for lexicography in many aspects. Tony McEnery, Richard Xiao and Yukio Tono state the advantages of using corpora in lexicography as follows. The greatest advantage of using corpora in lexicography lies in their machine-readable nature, which allows dictionary-makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds. The second advantage of the corpus-based approach, which is not available when using citation slips, is the frequency information and quantification of collocation which a corpus can readily provide. (McEnery et al, 2006)In fact, the *Collins Cobuild English Dictionary*, which is one of the corpus-based dictionaries, includes information about the frequencies of words, the detailed definitions of words, the usage in the authentic contexts and so on thanks to corpus-based analysis.

Although there are some corpora of South-East Asian languages which are able to be utilized as open resources, there are few studies concerning the dictionaries of these languages and the use of corpora. For example, in Thai, there are some open resource corpora, including the Thai National Corpus, Thai corpus in SEALANG LBRARY and ThaiWaC in Sketch engine, and there are a few studies about the entries in Thai-Japanese bilingual dictionaries.

Most of the Thai-Japanese bilingual dictionaries are general-purpose dictionaries, but they are used for many purposes, especially as an aid to language learning. And I believe that, among Thai-Japanese bilingual general-purpose dictionaries, the most famous dictionary is ‘タイ日大辞典’ 3rd edition compiled by Takejiro Tomita, who is an emeritus of Osaka University of Foreign Studies, because it has more information in each entry than other Thai-Japanese dictionaries, for example details of collocations, idioms, examples and proverbs, as well as cultural and historical information. In many cases, thus, it is also used as an aid to Thai learning for Japanese. Because of the recent politic, economic situations in Thailand, I

believe that the number of people who will study Thai will increase more in the near future. Therefore, with Thai corpus, I would like to search for some improvements in ‘タイ日大辞典’ 3rd edition in order to reinforce the side of an aid to Thai learning.

In this study, I focused on collocations comprised within two words because Thai has no inflection and changes of word forms, therefore it is useful to analyze and find the kinds and tendency of the collocations comprising two lexical words, which are similar to fixed expressions in Thai. I utilized the ThaiWaC corpus because it has the most data in presenting Thai corpora and is able to use statistical procedures such as Minimum sensitivity, Logdice and so on. From the function of Word List in ThaiWaC, I drew some lexical words of verbs, adjectives and nouns, and analyzed the collocations within two words related to them. And after getting the data of the tendency of them, I researched the authenticities of the collocations in‘タイ日大辞典’ 3rd edition with it.

Through this study, I aim to be able to present some improvements to ‘タイ日大辞典’ 3rd edition and provide some advice on selecting lexical words for Thai learners. I hope that this study will be devoted to the future development of ‘タイ日大辞典’.

References

Atkins, S. and Rundell, M. 2008. *The Oxford Guide to Practical lexicography*. Oxford University Press.

Béjoint, H. 1994. *Tradition and Innovation in Modern English Dictionaries*. Clarendon Press

Hartmann, R. R. K.. 2001. *Teaching and Researching Lexicography*. Pearson ESL

McEnery, T., Xiao, R and Tono, Y. 2006. *Corpus-Based Language Studies An Advanced Resource Book.* Routledge Applied Linguistics

Sinclair, J. 1987. *Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. Collins Cobuild

Svénsen, B. 1993. *Practical Lexicography Principles and Methods of Dictionary-Making*. Oxford University Press

Svénsen, B. 2009. *A Handbook of Lexicography The Theory and Practice of Dictionary-Making*. Cambridge University Press

Takejiro Tomita. 1997. タイ日大辞典 3rd edition. 日本タイクラブ

Towards a Multimodal Australian National Corpus

Michael Haugh[1], Steve Cassidy[2] and Pam Peters[3]

[1]*School of Languages and Linguistics*
Griffith University, Brisbane, Australia
[2]*Department of Computing*
Macquarie University, Sydney, Australia
[3]*Department of Linguistics*
Macquarie University, Sydney, Australia

Keywords**:** multimodal, speech, audiovisual media, annotation

Introduction

The Australian National Corpus has been established in an effort to make currently scattered and relatively inaccessible data available to researchers through an online portal. In contrast to traditional national corpora, it is conceptualised as an inter-linked database of multimodal and multilingual language resources that represents the Australian linguistic landscape, unified through common technical standards (Haugh, Burridge, Mulder and Peters, 2009). The inclusion of multimodal data, namely, digitised audio and audiovisual recordings alongside transcriptions or annotations, is driven in part by the movement into the hands of ordinary researchers of powerful technologies for digitizing and managing audio(visual) recordings, as well as transcribing or annotating such recordings, but also in part by the increasing interest amongst researchers working in sociolinguistics and pragmatics in the advantages of large-scale corpora (Baker, 2010; Dahlmann and Adolphs, 2009; Rühlemann, 2010). Corpora incorporating audio(visual) media have been used to investigate the role of backchannels (Knight and Adolphs, 2008; Wong and Peters, 2007), sociolinguistic variation in Australian Sign Language (Schembri and Johnston, 2007), and the interplay of gesture and intonation in dialogue (Mendoza-Denton and Jannedy, 2011), for instance. The inclusion of such media, however, raises a number of technical, ethical and legal challenges. In this paper, we first outline these challenges, and then discuss the legal-ethical and technical frameworks we have developed in order to incorporate such media into the Australian National Corpus. While acknowledging these challenges, we nevertheless argue that the development of a large-scale multimodal corpus is likely to raise new and rich avenues of research that extend the boundaries of corpus linguistics.

Technical, legal and ethical challenges

Each of the main collections that comprise the Australian National Corpus have different data formats, different systems and formats of annotation, and different approaches to metadata. This has raised challenges in terms of ingesting the data, annotations and metadata into common formats without losing valuable information. A myriad of complicated legal and ethical issues

have also arisen from the project as an outcome of making large amounts of language data from an array of sources available to other researchers and the public (the latter where permissible). These include the different forms of language data and issues of privacy to which each gives rise, the range of different approaches to ethical clearance (particularly in the case of legacy data collections), and the ownership of copyright versus moral rights residing in different individuals and/or institutions.

Technical framework

The technical architecture builds on the DADA system (Cassidy, 2010) and integrates separate data stores for the source media, meta-data and annotation behind a web based presentation layer based on the Plone content management system. All annotation in the corpus is stored as stand-off annotation, so the source media, be it text, audio or video, is stored separately in a web accessible location that is referenced by the meta-data and annotation stores. For audio and video resources this is standard practice; for the text based corpora this has meant generating markup-free versions of the text to act as the source media. Each 'document' (which could be text, audio or video) is given a unique URL by which it can be referenced or viewed directly.

The meta-data in the original collections is stored in a variety of formats ranging from spreadsheets to text embedded in tables at the start of Microsoft Word files. As part of the ingestion process, we parse this and generate a standard format that can be processed by XML tools. Meta-data is converted to RDF format and stored in a Sesame triple store. A significant challenge has been defining a common vocabulary that meta-data fields can be mapped to across the entire corpus. The design of this vocabulary has been mediated by the desire to remain compatible with existing standards such as Dublin Core and OLAC but also by a need to provide collection level meta-data to the Australian National Data Service (the funding body) in the RIFCS format. The resulting vocabulary is necessarily a hybrid of these different vocabularies and will expand as new collections are added that have their own unique meta-data fields.

A primary goal of the project is to bring together these varied collections under a common technical framework so that a rich set of end-user tools can be built to make them a more useful resource to the research community. In the first instance we are concentrating on providing a rich search and browse interface which makes use of the data, meta-data and annotation stored in the system.

Legal and ethical framework

It was initially decided that while attractive, the utilisation of Creative Commons licenses was not an acceptable option for the Australian National Corpus, because of the legal complexities that it would raise in the case of some of the collections. It was agreed that prospective collection owners of the Australian National Corpus should, at a minimum, allow the following usage rights to their collections: the means to use, adapt, publish, reproduce and communicate language data content for academic and research purposes through the Australian National Corpus discovery service. Additional copyright requirements regarding general terms and warranties were documented during both scheduled meetings and email correspondence. Drafts for a general Copyright and End User Licence agreements were produced with the help of members of the Australian National

Corpus legal group and inside legal counsel at Griffith University. The Copyright Licence drafting performed by Griffith University counsel was prepared for Griffith specific collections, with the intention of further modifications to be implemented by an outside counsel. The End User Agreement was drafted with reference to that used for the British National Corpus, and further modified by the legal group and outside counsel. Ethical clearance is treated as the responsibility of the collection owners, although they are required to adhere to Australian national standards for ethical behaviour in research.

References

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Cassidy, S. 2010 An RDF Realisation of LAF in the DADA Annotation Server. *Proceedings of ISA-5*, Hong Kong, January 2010.

Dahlmann, I.and Adolphs, S. 2009. Spoken corpus analysis: multimodal approaches to language description. In Baker, P. (Ed.), *Contemporary Corpus Linguistics*. London: Continuum, 125-139.

Knight, D. and Adophs, S. 2008. Multi-modal corpus pragmatics: the case of active listenership. In Romero-Trillo, J. (Ed.), *Pragmatics and Corpus Linguistics*. Berlin: Mouton de Gruyter, 175-190.

Mendoza-Denton, Norma & Stefanie Jannedy. 2011. Semiotic Layering through Gesture and Intonation: A Case Study of Complementary and Supplementary Multimodality in Political Speech. *Journal of English Linguistics*, 39 (3), 265-299.

Rühlemann, C. 2010. What can a corpus tell us about pragmatics? In O'Keeffe, A. and McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 288-301.

Schembri, A. and Johnston, T. 2007. Sociolinguistic Variation in the Use of Fingerspelling in Australian Sign Language: A Pilot Study. *Sign Language Studies*, 7(3).

A Study of Stance in Multi-type Phraseology

Anping He

School of Foreign Studies,
South China Norma University,Guangzhou, China,

Keywords**:** academic English stance, phraseology boundary, tendency, pedagogic discourse stance

Stance, as an important component expressing language users' purposes, intentions and attitudes, "is pervasive in all uses of language…and is a central topic for linguistics" (Stubbs, 1996:202). Yet it has until recently been largely neglected in foreign language teaching. This presentation is a project report on our study of academic stance from the perspective of multi-type phraseology. Phraseology is a multi-word basic unit of meaning in text or discourse. Its meaning is "formed and shaped by the persistent recurrence and coselection of words" (Sinclair, 2008:410) and can be retrieved automatically or through subjective identification. Its wording is not confined to adjacent connection but can "encompass all aspects of preferred sequencing" (Hunston, 2002:138). By widening the scope of phraseology study, researchers in corpus linguistics have opened up a "huge area of syntagmatic prospection" (Sinclair, 2004:19) and raised questions such as how to define boundary of a phraseology unit and what type of meaning is constructed by such unit.

Driven by these questions, we have tried to study academic English stance in terms of four types of co-selected word-sequences:

1) a prototype of epistemic stance "I think" with its propositional sequence;

2) colligation patterns of some modal-verb-type stance;

3) imperative sentence -- stem of stance;

4) 4-word lexical bundles automatically retrieved by computer means.

We take each as one type of phraseology unit and analyzed its recurrent pattern and meaning function in academic corpora that include spoken English corpus by university English learners and written corpus of EFL course-books in mainland China.

This presentation will demonstrate tendency of Chinese English learners' use of stance which is different from that of the native English speakers. It will also demonstrate changes of pedagogic discourse stance in EFL course-books during the current English curriculum reform. Discussion is made on the identification of "stances' phraseology boundary" and its strength in revealing nature of stance.

References

Hunston,S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Sinclair, J. 2004. Trust the text. In Sinclair, J. and Carter, R. (Eds.), *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. 2008. Envoi. In Granger, S and Meunier, F. (Eds.), *Phraseology: an Interdisciplinary Perspective.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

Can Concordance Improve the Lexico-grammatical Use of Abstract Nouns in L2 Writing?

Dora Zeping Huang

Department of English,
The Chinese University of Hong Kong

Keywords: concordance, corpus-informed, lexico-grammatical patterns, collocation, colligation

1. Introduction

In the past decade, a corpus approach has been regarded as a viable approach for helping learners with their lexico-grammatical patterns (e.g., Coxhead and Byrd, 2007; L. Flowerdew, 2010). By adopting a pretest-posttest experimental design, this study examines whether and the extent to which concordance has an impact on the lexico-grammatical use of abstract nouns in L2 writing.

2. Literature review (Omitted)

3. Research questions

Three research questions are formulated as follows:

1. Will concordances affect L2 learners' vocabulary use in their writing? If yes, in what ways does it affect learners' vocabulary use?
2. Will L2 learners think that concordances helped their vocabulary use in L2 writing?

4. Methodology

4.1 corpus-informed materials

The topic-specific corpus was established at the very beginning. It was composed of texts related to the writing topics concerning "gambling" and "lottery". The texts were obtained from two sources. One was from the online authoritative English news websites; the other was from a ready-made small corpus *LOCNESS* (Louvain Corpus of Native English Essays).

Next, a keyword list of the topic-specific corpus was generated with the aid of a corpus tool Wmatrix (Rayson, 2002). Target words were selected from the keyword list on the basis of three criteria (Read, 2004): 1) frequency occurring in the topic-based small corpus (each word occurs at least three times); 2) abstract nouns often used in opinion essays. According to these two criteria, five words were chosen. They were *controversy, criticism, objection, situation* and *effect*. Subsequently, about ten concordance lines of each target word were selected and presented in the corpus-informed materials.

4.2. Experimental design

Forty Chinese EFL learners were randomly assigned to the control group and the experimental group. At the pre-writing stage, both groups were given the five target abstract nouns. The experimental group was supplied with the corpus-informed concordances while the control group consulted dictionaries about the usage of the words. The written texts of the pretest, immediate posttest and delayed posttest were analyzed and compared between and within groups. Table 1 shows the procedures of the experiment.

Table 1: Procedures of the experiment

Participants	Week 1 Pretest	Week 2 Immediate Posttest	Week 4 Delayed posttest	Week 5 Survey
Experimental group (20 students)	-Brainstorm (5 mins); -Write an essay on "the impact of the tourist industry") (60 mins)	-Study the concordance exercises (15 mins) and brainstorm (5 mins); -Write an essay on "lottery" (60 mins)	-Brainstorm (5 mins) -Write an essay on "gambling" (60 mins)	-Do a questionnaire on evaluating concordance exercises (20 mins)
Control group (20 students)	Same as above	-Consult the given words in dictionaries if necessary (15 mins) and brainstorm (5 mins); -Write an essay (60 mins)	Same as above	NA

5. Results

The quantitative and qualitative results showed that the experimental group, as compared with the control group, used the five target nouns with a higher variety of collocational and colligational patterns in the immediate posttest and delayed posttest. In addition, fewer linguistic errors in using these nouns were found in the experimental group than the control group. The post-experiment learning journals and questionnaires administered to the experimental group further confirmed that concordance exercises encouraged usage-based learning, helped students notice the lexical collocations and prepositional colligations of the target words, and thus improved accuracy and complexity in their productive language. Despite these positive findings, potential problems of using concordance materials for independent learning were also reflected in the students' written output and reported in the learning journals.

References

Coxhead, A. and Byrd, P. 2007. Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, *16*, 129-147.

Flowerdew, L. 2010. Using corpora for writing instruction. In McCarthy, M. and O'Keeffe, A. (Eds.), *The Routledge handbook of corpus linguistics. Lo*ndon: Routledge, 444-457.

Rayson, P. 2003. Matrix: A statistical method and software tool for linguistic analysis

through corpus comparison. PhD. thesis, Lancaster University.

Read, J. 2004. Research in teaching vocabulary. *Annual review of Applied Linguistics*, 24, 146-161.

Criterial Features of L2 Acquisition Stages: Contrastive Interlanguage Analysis at Elementary Level

Taku Kaneta
Graduate School of Area and Cultural Studies,
Tokyo University of Foreign Studies, Tokyo, Japan.

This study aims to identify L2 learners' criterial features which characterize their proficiency levels. Since the release of the International Corpus of Learner English (ICLE) (Granger, Dagneaux, and Meunier, 2002), learner corpus research has been in the ascendant, illuminating various interesting characteristics of interlanguage, such as underuse vs. overuse phenomena, transitional error patterns, and developmental patterns in multiple aspects of lexico-grammatical features (Granger, Hung, and Petch-Tyson, 2002). One of the recent research interests of the field is to identify criterial features, i.e. linguistic features which can serve as criteria indicating particular proficiency level of language learners (Tono, 2010). The present study attempts to contribute to this trend by identifying criterial features focusing on elementary level learners with different L1 backgrounds.

The study used the International Corpus of Crosslinguistic Interlanguage (ICCI), a corpus of written compositions by more than 6,000 learners of English ranging from the third-year elementary school to the third-year senior high school (Tono, 2009). The learners were sampled from 7 different countries/regions, which enabled contrastive interlanguage analysis in this study. The corpus was tagged by POS and linguistic features reported to function as criteria features (North, Ortega, and Sheehan, 2011), to carry out multivariate statistical analyses (e.g. correspondence analysis). The research focused on the elementary level in particular, sifting out lexico-grammatical properties in relation to different developmental stages (the school year), together with L1-specific features.

The analysis is still to be completed due to the data recheck in process and will be presented at the conference. Comparison with similar research such as Hawkins and Buttery (2010) or Tono (2000) will be made to discuss implications for future research.

References

Granger, S., Dagneaux, E. and Meunier, F. 2002. The International Corpus of Learner English. Handbook and CD-ROM. Version 1.1. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Hung, J. and Petch-Tyson, S. 2002. *Computer Learner Corpora*. Amsterdam: John Benjamins.

Hawkins, J. A. and Buttery, P. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal, 1*(01).

North, B., Ortega, A. and Sheehan, S. 2011. *Core inventory for general English*: British Council

EAQUALS.

Tono, Y. 2000. *A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora*. Paper presented at the PALC'99: Practical Applications in Language Corpora.

Tono, Y. 2009. Variability and Invariability in Learner Language: A Corpus-based Approach. In Yuji, K., Makoto, M.and Jacque, D. (Eds.), *Corpus analysis and variation in linguistics,* Vol. 1. John Benjamins Pub Co., 67-82.

Tono, Y. 2010. Learner Corpus Research: Some Recent Trends. In Weir, G.. and Ishikawa, S. (Eds.), *Corpus, ICT, and Language Education.* Glasgow: University of Strathclyde Publishing, 7-17.

Gerund and Infinitive after *Like*:

An Investigation of *Like to Say* and *Like Saying*

Heng-ming Carlos Kang

Graduate Institute of Linguistics
National Chengchi University,
Taipei, Taiwan

Keywords**:** gerund, infinitive, verbal complement, verbs of liking, verbs of saying

Introduction

In English grammar, some verbs must collocate with a gerundive complement (e.g. *enjoy singing*), some verbs must collocate with an infinitive complement (e.g. *want to sing*). There are still some verbs that take both gerundive and infinitive complements, such as *like*, *love*, *start*, and *try*. This type of verb has been of interest to linguists. Many linguists have been puzzled about the difference between the gerundive and infinitive complements when they follow the same verb.

Previous approaches

Woods (1956) treated the difference between gerunds and infinitives with generality, claiming that *Lying is wrong* had a universal sense, while *To lie is wrong* had a specific sense.

Givón (2001) and Dixon (1995), on the other hand, demonstrated that the gerund complement tends to denote that the action is performed, already in existence, whereas the infinitive complement tends to denote that the action is not performed. Example (1) from Givón (2001) shows this difference. In (1a), John changed his mind after doing the dishes every morning; in (1b), John changed his mind before he did the dishes every morning.

(1) a. John started doing the dishes every evening, but then changed his mind.

b. John started to do the dishes every evening, but then changed his mind.

Duffley (2004), on the other hand, adopted Bladon's (1968) framework and claimed that the difference is enjoyment and desire as in (2). In (2a), the gerundive complement denoted enjoyment while the infinitival complement in (2b) denoted desire.

(2) a. He likes going home by car.

b. He likes to go home by car.

The disadvantage of the previous approaches lies in that they all used invented sentences (except for Duffley (2004)) as examples. Duffley (2004), despite using corpus sentences to demonstrate his ideas, cannot identify clearly whether *like going home by car* is enjoyment rather than desire since the idea of enjoyment and desire is a psychological concept which is determined by the speaker himself.

Givón's temporal approach was adopted in this study. The purpose of this study is to

investigate the temporal difference between gerundive and infinitival complements for verbs of liking. With the corpus data, the real language use can be captured. In addition, the context can provide indirect evidence of the temporal reference of the actions.

Corpus and the investigated pattern

The data came from The Corpus of Contemporary American English (COCA) (Davies, 2008-). In order to give a detailed and close examination of the verbal complements, the scope of this study was narrowed down to the particular constructions *like to say* vs. *like saying*. Based on the discoursal position of the said content, the syntactic arrangements were categorized into four types: before-pattern, after-pattern, both-pattern and separated-pattern. (XXX for the said content. The quotation marks were considered optional.)

(3)

a. before-pattern: "XXX"…like to say/ saying.
b. after-pattern: …like to say/ saying "XXX."
c. both-pattern : "XXX"……like to say/ saying "XXX."
d. separated-pattern: "X" … like to say/ saying "XX."

In addition to the discoursal position, the syntactic forms of the said content demonstrated seven types: noun phrase, verb phrase, adjective phrase, preposition phrase, anaphoric pronoun, cataphoric pronoun and sentence.

It was hypothesized that, based on Givón's idea of temporality (2001), if the gerund complement tends to denote a performed action, it is likely to collocate with the both-pattern, because the both-pattern designates that what is going to be said has been said before. On the other hand, if the infinitive complement tends to denote an action to be performed, it is likely to collocate with the after-pattern. The second hypothesis was that if the gerund complements tend to denote performed action, anaphoric pronouns will tend to collocate with *like saying* rather than *like to say*.

Results

With the help of query syntax, the corpus returned 263 tokens for like saying and 1411 tokens for *like to say*. A random sampling of 300 was done for the 1411 tokens of *like to say*. For the 263 tokens of *like saying* and 300 tokens of *like to say*, further manual examination was done to eliminate invalid constructions. Final valid number of tokens is 33 for *like saying* and 218 for *like to say*.

The major syntactic arrangement *like saying* adopts is both-pattern (72.73%), which only accounts for 5.96% in *like to say*. On the other hand, the major pattern collocated with *like to say* is the after-pattern (63.76%). The results are consistent with the temporality hypothesis.

As for the syntactic form, the anaphoric pronoun collocated with *like saying* most (60.61%), and only occurs in both-pattern. This is consistent with the temporality hypothesis. On the other hand, anaphoric pronoun accounts for very little in *like to say* (only 3.79%). Instead, sentences tend to collocate with like to say (78.67%). In addition, more forms can collocate with *like to say* than *like saying*.

Acknowledgements

I am very grateful to Doctor Siaw-Fong Chung for her suggestions in the improvement of the paper. I would also like to thank my classmates in the course Corpus Linguistics and Language Teaching for their comments on an earlier version of this paper. I am solely responsible for any errors and inadequacies in the present final version.

References

Bladon, R. A. 1968. Selecting the to- or -ing nominal after like, love, hate, dislike, and prefer. *English Studies*, 49, 203-214.

Davies, Mark. 2008-. The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. Available online at http://www.americancorpus.org.

Dixon, R. M. W. 1995. Complement clauses and complementation strategies. In Palmer, F. R. (Ed.), *Grammar and meaning: Essays in honour of sir john lyons.* England: Cambridge U Press, 175-220.

Duffley, P. J. 2004. Verbs of liking with the infinitive and the gerund. *English Studies*, 85(4), 358-380.

Givón, T. 2001. *Syntax: An introduction, II*. Amsterdam: Benjamins.

Wood, Frederick T. 1956. Gerund versus infinitive. *English Language Teaching* 11, 11-16.

Developing and Analyzing Corpora from Subtitles

Kenji Kitao[1] and S. Kathleen Kitao[2]

[1]*Department of Culture and Information Science*
Doshisha University, Kyoto, Japan
[2]*Department of English*
Doshisha Women's College, Kyoto, Japan

Keywords**:** subtitles, speech acts, using movies in ELT

Introduction

Movies and television programs can be useful in teaching second/foreign language students about using language for communication. They provide a context, information about relationships among interlocutors, nonverbal information, etc. One way that they can be used is in analyzing dialogue to learn about speech acts (greetings, requests, invitations, etc.) and show how different expressions are used in different contexts, ways in which interlocutors are polite or impolite to each other, etc. They can also be used to look at how language is used differently in different contexts, between people of different status, and so on. In some cases, examples from movies or television programs can be helpful in illustrating concepts from linguistics or communication, particularly lingustic pragmatics.

Getting Subtitles

Teachers or students who want to use dialogue from movies or television programs can access subtitles in one of two ways – downloading subtitles from DVDs or finding transcripts or scripts online.

One of the programs that can be used to download subtitles from DVDs is DVD Decrypter, and it can be downloaded at http://www.mrbass.org/dvdrip/. Another is SubRip, which can be downloaded at http://www.divx-digest.com/software/subrip.html. If you use programs that download subtitles, you should be aware that the subtitles are not always exactly like the actual dialogue. However, the differences are not usually significant. If it is important for you to have the exact wording from the dialogue, you can use these programs to download the subtitles and then check them against the dialogue, making changes as necessary.

There are also a number of websites where transcripts or scripts of movies are posted. The following are four such sites:

Drew's Script-o-Rama http://www.script-o-rama.com/oldindex.shtml

Movie Scripts and Screenplayshttp://www.moviescriptsandscreenplays.com/

Scriptshttp://www.movie-page.com/movie_scripts.htm

Simply Scriptshttp://www.simplyscripts.com/

Allsubs.orghttp://www.allsubs.org/

Some scripts are early drafts, and they may differ considerably from the dialogue in the movie. Transcripts may also contain errors, so they, too, should be checked against the movie itself.

Subtitles can be copies and pasted into Excel, with one utterance on each line, or one turn on each line, depending on how the teacher intends to use them. Information can be added, including the name of the speaker, the sex of the speaker, and the function.

Teaching Using Subitiles

Subtitles can be used as teaching tools or in the classroom for data driven learning. The words and sentences in the subtitles can be analyzed in various ways. You can look at the average number of words per sentence, the level of difficulty of the vocabulary (for example, a program at http://www.tcp-ip.or.jp/~shim/j8web/j8web.cgi analyzes the level of frequency of the words in a text). Also, you can use Key Words in Context programs to look at how particular words are used. (The Antconc concordancing program, which is available as freeware, can be found at http://www.antlab.sci.waseda.ac.jp/ software.html.) You can look at the speech acts in the movie and compare how, for example, requests or apologies are used in different situations. You can compare how younger or older speakers use expressions, or how male and female characters use them differently.

For example, we used software to scan the subtitles from the 1998 movie *You've Got Mail*, which we chose because we thought it was a movie that students would enjoy, because it has a variety of situations with natural, everyday dialogue, because it involves a variety of speech acts in a variety of situations. We entered the subtitles into Excel, with one line for each utterance. We used different columns to identify each speaker, the sex of the speaker, and the function of each utterance. Then we looked at how different request expressions were used. We also analyzed one scene with a series of requests by two different people in terms of Brown and Levinson's theories of positive politeness (that is, satisfying the hearers need to be liked and feel included by using in-group markers, showing interest in the hearer, making jokes, etc.) and negative politeness (that is, showing respect for the hearer by not assuming that the request will be accepted, minimizing the size of the request, apologizing, etc.). This scene has been used in linguistics pragmatics courses to help students understand the concept of positive and negative politeness. In class, Brown and Levinson's theory and categories of politeness are discussed, and then students watch the scene and receive a transcript of it. They work in groups or pairs to find examples of positive and negative politeness and identify the specific strategies used. Their results are then discussed with the class. (The scene can be found at http://www.youtube.com/watch?v=q3tRKs9o_2A.)

Conclusion

Because of the context that they provide, movies are useful for English language teaching, and particularly for the teaching of speech acts. Downloading subtitles from DVDs or finding scripts or transcritps online makes it possible to analyze usage of words or speech acts. Students can study speech acts through data driven learning, and examples can be found to use as example of concepts from lingustics or communication.

A Corpus-Based Study of University Entrance Exams

S. Kathleen Kitao[1] and Kenji Kitao[2]

[1]*Department of English*
Doshisha Women's College, Kyoto, Japan
[2]*Department of Culture and Information Science*
Doshisha University, Kyoto, Japan

Keywords**:** entrance exams, vocabulary, readability

Introduction

In Japan, written exams are central to the university entrance system, and English language exams can make a significant part of the total exam score. Individual universities have their own exams, and there is a national exam, called the Center Exam, which is used either alone or with interviews or with another written exam. The largest part of these exams is usually made up of reading passages with multiple choice questions or cloze-type items. They also include grammar and vocabulary questions, translation questions, and so on

The validity and reliability of entrance exams are influenced by a number of factors, but appropriate lexical difficulty and readability for reading passages are very important. These are not sufficient by themselves to make valid and reliable tests, but they provide the foundation of the exam. Therefore, it is useful to look at lexical difficulty of tests and the readability of reading passages to evaluate that aspect of their validity.

In two previous studies (Kitao and Kitao, 2006; Kitao and Kitao, 2009), we analyzed the lexical difficulty and readability of reading passages in the 2005 and 2008 entrance exams of four private Japanese universities and the Center Exam over a period of 16 years from 1990 to 2005. In this study we will compare those results with the 2010 exams of the same universities and the national exam.

Research Questions

We considered two research questions.

1. What is the level of difficulty of words used in the English entrance exams of four major private universities in 2010, and how has this changed between the 2005 and 2008 exams?
2. What is the level of readability of reading passages used to test reading comprehension, and how has this changed since the 2005 and 2008 exams?

Methods

We analyzed the English sections of the entrance exams of four private Japanese universities using word frequency counters and measures of readability. We included exams from four private

universities. (Each university gives more several English exams to students with different majors or on different days.) For University A, we analyzed 7 exam; for University B, 8 exams; for University C, 9 exams; and for University D, 8 exams. Each individual test and all of the tests were analyzed for each university.

In analyzing the vocabulary frequency, we used every word in English in the exams, including reading passages and questions but removed Japanese instructions, question numbers and letters, and so on. We used the Japan Association of College English Teachers 8000 (JACET) word list, which was developed based on the British National Corpus, but adjusted to reflect language education in Japan. It has 8 levels of 1000 words each. Using a web-based program (http://www.tcp-ip.or.jp/~shim/j8web /j8web.cgi), we calculated the percentage of total words and unique words at each level. We looked at the percentage of words at each of the eight levels, and words not included in the 8 levels but excluded contractions, proper nouns, and non-words from the analysis. We used the results to calculate the number percentage of words over the 4000-word level for total words and unique words. According to Barrow, Nakanishi, and Ishino (1999), the average Japanese college freshman has a vocabulary of 2304 English words. Private university exams are probably taken by better-than-average students, so we used the 4000-word level for our evaluation (that is, levels 1-4).

For readability analysis only reading passages, excluding dialogues, were used. For cloze passages, we filled in the blanks and used the intact passages. Software was used to calculate two readability scales, the Flesch Reading Ease (FRE) scale and the Flesch-Kinkaid Grade Level (Flesch-Kinkaid Grade Level) as well as other statistics, including the number of words in reading passages in each test and the number of words per sentence. Readability scales take into account grammatical complexity (represented by sentence length) and lexical difficulty (represented by the number of letters or syllables in words). The FKGR is represents readability in terms of US grade levels. A typical American newspaper story, which would be fairly difficult for most Japanese university students to read, is around 8 on the FKGL, meaning that it could be read by US students in the 8th grade or higher. On the FRE, the higher the score, the easier the reading passage would be. For native English speakers, a FRE score between 30 and 49 indicates a very difficult reading, and 50-59 is fairly difficult. While the readability scales were developed for native English speakers, Greenfield (2004) found high correlations between the two scales and measures of results of cloze tests with Japanese university students. This indicates that these readability scales reflect the relative difficulty of reading passages for Japanese readers.

Results and Discussion

The level of difficulty of the vocabulary of the reading passages on the university entrance exams considered in this study would be difficult for most students to handle, and this situation has not improved since studies of the entrance exams for 2005 and 2008. In 2010, all four universities had at least some tests with over 5% unknown words (i.e., words less frequent than the first 4000 words on the JACET list), and for two of those universities, the average for all the tests was over 5% of unknown words. The Center Exam had less than 5% unknown words. One of the universities had an average level of readability at the 12th grade level, and two at the 10th grade level. These levels of readability would mean that the passages would be very difficult for Japanese students These results could indicate that there is a threat to the validity of these tests,

since incomprehensible reading passages or a high level of unknown lexical items would lead students to guess and to use test-taking strategies to try to figure out the answers rather than depending on an understanding of the reading passages and other types of items to choose their answers.

The JACET 8000 list is useful as a standard to check the level of difficulty of vocabulary. There are programs available online for making use of this frequency list. One (http://www01.tcp-ip.or.jp/~shin/j8web/ j8web.cgi) shows the percentage of words at each level, both in terms of total words and unique words. It also provides a list of the words at each level, with the number of occurrences. In addition, at the JACET 8000 Level Marker (http://www01. tcp-ip.or.jp/~shin/J8LevelMarker/j8lm.cgi), reading passages can be entered, and the software produces a text with words color coded according to their level of difficulty.

In addition, online calculators of readability can be used to check the readability of reading passages. For example, the one available at http:// www.online-utility.org/english/readability_ test_and_improve.jsp calculates various statistics about a reading passage, such as the number of words per sentences, as well as several readability scales. It also lists sentences that might be problematic. This would make it easy for test makers to identify difficult sentences.

References

Barrow, J., Nakanishi, Y. and Ishino, H. 1999. Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, 27, 223-247.

Chujo, K. n.d. Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. Retrieved September 30, 2010 from http://www5d. biglobe.ne.jp/ ~chujo/eng/data/ rodopi.pdf.

Greenfield, J. 2004. Readability formulas for EFL. *JALT Journal*, 26(1). Retrieved September 30, 2010 from http://www.jalt-publications.org/jj/articles/ 2004/05/greenfield.

Kitao, K. and Kitao, S.K. 2006, August 19. Corpus-based analysis of Japanese university entrance exams. Paper presented at The 2006 Asia TEFL International Conference, Hakata, Japan. Retrieved September 30, 2010, from http://www.cis.doshisha.ac.jp/kkitao/library/handout/ 2006/ASIA/corpus-based.pdf.

Kitao, S. K. and Kitao, K. 2009. An analysis of Japanese university English entrance exams using corpus-based tools. In Koyama, T., Noguchi, J., Yoshinari, Y., and Iwasaki, A. (Eds.), *Proceedings of the WorldCALL 2008 Conference*. Osaka: The Japan Association for Language Teaching and Technology, 72-75. Retrieved September 30, 2010 from http://www.j-let.org/~wcf/modules/tinyd12/

New Learner Corpus Functionality in the Sketch Engine

Vojtěch Kovář[1,2] and Diana McCarthy[2]

[1]NLP Centre
Faculty of Informatics, Masaryk University, Brno
[2]Lexical Computing Ltd., UK

Keywords**:** Sketch Engine, learner corpus, concordance, error tagging

Introduction

We present an interface for efficient searching in learner corpora with annotated errors that is built into the corpus querying tool Sketch Engine (Kilgarriff et. al., 2004).

We describe the data format needed for an annotated learner corpus, which allows for nested errors of different types as well as correction mark-up. Then, we show a specialized web user interface designed for searching learner error corpora.

The Sketch Engine

The Sketch Engine is a powerful corpus-querying tool used by lexicographers, language teachers and language learners throughout the world. Its main functions are:

- a concordancer that allows querying in an extended CQL (corpus query language) syntax (Jakubíček et al., 2010) and provides also a simplified and intuitive querying interface for users that are not familiar with CQL
- a word list feature that allows listing of words from the corpus according to various criteria, including extracting keywords from a sub-corpus and also allows for frequency lists of meta-data (such as domain of the documents)
- a word sketch feature that provides one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour
- a statistical thesaurus based on the collocations given by the word sketches

All of these features are available as a web service at *http://www.sketchengine.co.uk*, together with preloaded corpora for more than 50 languages and a system that allows building user corpora.

Learner Corpora

Based on a customer requirement, a new functionality has been added into Sketch Engine that facilitates work with learner corpora marked for errors and corrections. Together with the newly implemented features, learner corpora can be very useful for language teachers and authors of course books to discover common types of errors and characteristic patterns of text where people tend to make errors. In the following, we will illustrate the usage of error corpora within the system.

Source Data

Usually, the Sketch Engine system require sa vertical format (word-per-line with possibly more

columns for lemmas, tags etc. and the structure XML-like mark-up) as the source data format. The errors are marked up using <*err*> and <*corr*> tags for errors and their corrections respectively. The type of error can be specified with the *type* attribute of either of the tags. An example of a small source text is shown in Figure 1.

Searching possibilities

The display of the error mark-up in the concordance lines can be highly customized – the system can display the full mark-up, as in the source vertical file, or an abbreviation for readability. In Figure 2, we show what the concordance lines can look like for the data from Figure 1. Using the frequency function, users are able to obtain the frequency distribution of e.g. error types for a given query. From this result, the concordance can be filtered for a particular error type. It is also possible to get the error type frequency distribution for the whole corpus or a sub-corpus.

We have developed a special query interface for the users that translates the commonest queries for learner corpora into CQL (words within the error, words within the correction or type of the error – see Figure 3) so that the users do not have to use the CQL themselves.

Error mark-up can be exploited in the Word Sketch or Thesaurus functions to produce word sketches specific for error mark-up. We hope to explore this in future work.

```
We          we          PR
learn       learn       VV
maths       math        NN
to          to          PP
<err type="BadWording">
<err type="Typo">
caan        caan        ??
</err>
<corr type="Typo">
can         can         VA
</corr>
</err>
<corr type="BadWording">
be          be          VB
able        able        VP
to          to          PP
</corr>
compute     compute     VV
our         our         PR
taxes       tax         NN
```

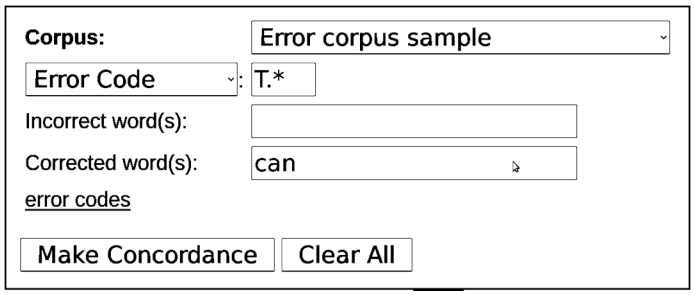


Figure 1. Example of the source learner data. **Figure 3.** Error query interface

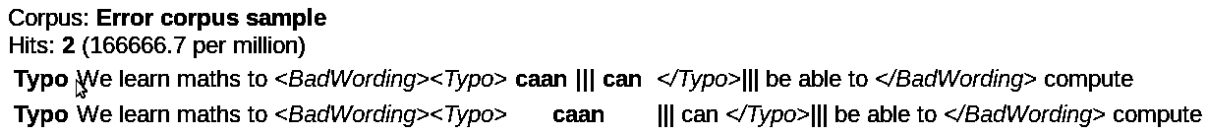


Figure 2. One possible type of display of the error data in the concordance. The first line is for the "Bad Wording" error, the second is the "Typo" error.

Conclusion

We have introduced a powerful tool for searching learner corpora with annotated errors. We described its potential and gave examples of its use. We hope the tool will prove useful for those working with learner corpora, language teachers and authors of the learner books.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536, in the National Research Programme II project 2C06009, by the Czech Science Foundation under the project P401/10/0792 and by the EU project PRESEMT (ICT-248307). We would like to thank Cambridge University Press for their collaboration in this development.

References

Jakubíček, M., Rychlý, P., Kilgarriff, A. and McCarthy, D. 2010. Fast syntactic searching in very large corpora for many languages. *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Waseda University, 741-747.

Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. 2004. The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress*. Oxford : Oxford University Press.

Using a Web Corpus in Indonesian Language Learning

Deny A. Kwary[1] and Adam Kilgarriff[2]

[1]*Faculty of Humanities, Airlangga University, Indonesia*
[2]*Lexical Computing Ltd.*

Keywords: Indonesian language, Sketch Engine, Google, language learning

1. Introduction

Indonesian is a new language, adopted as the national language on independence on 17 August 1945. However it is not widely used as a mother tongue and most Indonesian people have problems using it correctly. The government is working to improve the level of proficiency. A resource that would support this goal is a corpus. Until recently, there has been no large, general Indonesian corpus available. Lexical Computing Ltd has now developed IndonesianWaC and made it available in the Sketch Engine.

The question of whether a corpus is good or not, depends on what we want to use it for (Kilgarriff et al. 2010). In this paper, we describe how the corpus can be used to support Indonesian language learning. In particular, we suggest the creation of a High Frequency Wordlist of the Indonesian language, and the use of corpus examples, collocations and thesaurus entries to make a new dictionary and grammar. We also compare results from the corpus with the standards set by the Centre for Language Cultivation and Development (*Pusat Pembinaan dan Pengembangan Bahasa*), and results of Google searching to compare how the language has been prescribed with how it has been used. Finally, we make some proposals for further improving the quality of the corpus.

2. Overview of the Sketch Engine and the IndonesianWAC

The Sketch Engine is a corpus query system (CQS) which, in addition to general CQS functions, provides 'word sketches', one-page, automatically-generated summaries of a word's grammatical and collocational behaviour, and also an automatic thesaurus, and 'sketch differences', identifying the similarities and differences between the collocational behaviour of two similar words (Kilgarriff et al. 2004). The Sketch Engine currently offers corpora for forty languages. IndonesianWaC contains approximately 100 million Indonesian words (see Kilgarriff et al. 2010 for its method of construction).

Some of the corpora in the Sketch Engine have been POS–tagged, which supports word sketches and other advanced functionality. IndonesianWaC has not been POS-tagged. However, we can still use the other menus at the Sketch Engine with this Indonesian corpus.

3. The Situation of Indonesian Language Use

Most Indonesian people have problems using the national language correctly. This may be due to the fact that there are not many Indonesian people whose mother tongue is Indonesian, and the great influence of foreign languages, especially English, in Indonesia. Alwi et al. (1998: 1) find

that the number of Indonesian speakers is less than that of local languages, such as Javanese and Sundanese. In addition, many schools in Indonesia use English as a medium of instruction in order to get the title as international standard schools.

In response to these concerns, the Government issued Law No. 24/2009 which prescribes the use of the Indonesian language in national forums. Then, in 2010 the Centre for Language Cultivation and Development published *Panduan EYD dan Tata Bahasa Indonesia* ('Guidelines on Indonesian Standard Spelling and Grammar'), which contains lists of standard vs. non-standard Indonesian words, phrases and grammatical constructions. However, we find that some of these prescribed standard forms contradict to the actual language use. Therefore, in the following section, we suggest the use of a descriptive approach, using the Indonesian corpus to support Indonesian language.

4. Using the Corpus to support Indonesian language learning

The effectiveness of using high frequency words for language learning and for creating dictionaries have been proven in many cases, especially for the English language learning. Therefore, we suggest that the first step which should be taken in order to create an Indonesian learner dictionary and a grammar book is creating a High Frequency Wordlist of the Indonesian language. This can be done by using the Sketch Engine to automatically create a frequency list of all Indonesian words in the corpus. Before selecting the 1,000 most frequent words from the list, we need to clean the list. This process may include deleting words which are not Indonesian, correcting and combining the frequencies of a word which has two or more variant spellings, and checking against systematic relations among words in order to avoid omissions.

When the cleanup process has been completed, the wordlist can then be used to determine the lists of most frequent words. Consequently, it will be possible to create graded grammar books based on this frequency list. The high frequency words can also be used as the defining vocabulary for an Indonesian learner dictionary. Up to now, there are no Indonesian grammar books and dictionaries which are based on a corpus. Efforts towards creating corpus based grammar books and dictionaries should be initiated.

Using the corpus may mean that some of the so-called standard forms established by the Government to be changed or reconsidered. When we used the corpus to compare the frequencies of Standard vs. Non-standard forms listed in the Guidelines on Indonesian Standard Spelling and Grammar, we find favourable results for most Standard forms. For example, the frequencies of *bertemu* (Standard form) and *'ketemu*' (Non-standard form), both meaning 'to meet', are 14,265 and 2,109, respectively. However, the frequency of the Standard form *cenderamata* 'souvenir' is lower than that of the non-standard form *cinderamata*, i.e. 106 and 145, respectively. When we crosschecked the results with the Google.co.id setting the language to Indonesian only, the results were 547,000 for *cenderamata* and 1,600,000 for *cinderamata* (search results on 14 October 2011). In this case, we suggest that the spelling *cinderamata* be considered the standard form.

If, however, a non-standard form or construction only has a higher frequency than the standard one in the Google search, but not in the corpus, we are inclined to consider it a standard form. For example, the comparison between *itu benar* and *itu adalah benar* 'that is correct'. The results of the Google search on 14 October 2011 shows 4,440,000 hits for *itu benar* and 6,120,000 hits for *itu adalah benar*. However, the frequencies of *itu benar* and *itu adalah benar* in the corpus are 1,375 and 114, respectively. Consequently, we suggest that a corpus-based grammar

book should still consider *itu benar* as the standard construction. In addition to the Wordlist menu in the Sketch Engine, the collocations and thesaurus menus available at the Sketch Engine are also useful in creating corpus-based Indonesian language learning resources.

5. Suggestions for Augmenting the Corpus

The quality of IndonesianWaC could be augmented by further cleaning, and POS-tagging. By 'cleaning' we mean further processes of removing duplicates, spam and other non-natural text, and material in the wrong language (most often, the closely-related Malay). After the cleaning process is completed, it is necessary to add the POS tags, so that we can use the Word sketch and Sketch difference menus. With the Word sketch, we will be able to see the grammatical constructions of a particular word, and with the Sketch difference, we will be able to determine the similarities and differences between two synonyms. These will enable Indonesian authors to create better Indonesian grammar books and dictionaries which can be used as better Indonesian language resources, and to use the corpus for Indonesian language learning.

References

Alwi, Hasan, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 1998. *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka.

Kilgarriff, A. Rychly, P., Smrz, P., and Tugwell, D. 2004. "The Sketch Engine". *The Proceedings of the 11th Euralex International Congress*. Lorient, France.

Kilgarriff, A., Reddy A., Pomikalek, J. and Avinesh, PSV. 2010. A Corpus Factory for Many Languages. *The seventh international conference on Language Resources and Evaluation*. Malta: ELRA

Ministry of Laws and Human Rights, Republic of Indonesia, Laws *No. 24/2009*, on 9 July 2009.

Pusat Pembinaan dan Pengembangan Bahasa. 2010. *Panduan EYD dan Tata Bahasa Indonesia*. Jakarta: TransMedia.

Analysing Story Structure with a Bilingual Corpus of Aesop's Fables

Oi Yee Kwong

Department of Chinese, Translation and Linguistics
City University of Hong Kong

Keywords**:** story structure, discourse annotation, Aesop's Fables, bilingual corpus

Introduction

As a special type of discourse, stories have their own unique structure which distinguishes them from other genres of discourse. Corpus-based analyses of different subtypes of stories (e.g. nursery tales, short stories, etc.) under different frameworks have been done in the past (e.g. Hasan, 1996; Toolan, 2009). Fables, as a special kind of stories, share many of the properties of stories in general, but at the same time have their own characteristics, such as they are usually associated with a moral. Thus structurally they might resemble simple stories, presenting the events happening to a protagonist in temporal sequence, but semantically they often go beyond the surface meanings of the text to teach a lesson. Stories in Aesop's Fables, for instance, are relatively short originally, but the concise stories can somehow effectively convey the morals. Hence fables could be structurally simple on the one hand, but often semantically deep on the other, and it is therefore more cognitively demanding to understand fables. This may account for the way in which fables are retold to children: usually at greater lengths, with more episodes and more explicit descriptions of the deeds and thoughts of the protagonists, to guide the younger readers through the lesson. How these longer versions achieve this purpose, through different linguistic means, is what we are interested to find out in this study.

A Bilingual Story Corpus

To provide a useful resource for research on story analysis and story understanding, a bilingual corpus containing various published versions of Aesop's Fables in English and Chinese is compiled. The corpus is annotated at the structural level for the role of individual discourse segments and how they group together into larger structural units, and the semantic level for how discourse segments relate to one another semantically to form a coherent text. In designing the annotation schemes, we have drawn on previous analyses on stories and discourse, and considered annotator-friendliness.

For structural annotation, we started with previous studies on story grammar (e.g. Rumelhart, 1975; Mandler and Johnson, 1977), despite the controversy on its appropriateness for describing story structure (e.g. Wilensky, 1982). We observed that many of the constituent labels in Rumelhart's (1975) story schema, for instance, are more semantic in nature, e.g. Reaction, Goal Path, etc. In fact, basically only the high level nodes like Episode can really be considered "structural", corresponding to the grouping of individual discourse segments into larger chunks. The terminal nodes, State and Event, can nevertheless be suggested by the lexical items (particularly verbs) in individual discourse segments, but others like the many intermediate level nodes require a lot more inference and sometimes subjective judgement. So

we restrict the structural level tags for our annotation to Setting, Episode, Internal Event, Internal State, Event, State, Speech, and Moral. Constituents involving semantic interpretation are incorporated in the semantic annotation instead.

The most important element regarding the semantics of fables has to do with the expression of the moral, and this is often closely associated with the intention of the writer. The intentional structure in Grosz and Sidner's (1986) discourse model, for instance, should be able to account for this aspect, but it is somehow underspecified for actual annotation. It would be more feasible to ask annotators to identify discourse relations most relevant to fables to start with. With reference to Rhetorical Structure Theory (Mann and Thompson, 1987) and the Penn Discourse Treebank (Miltsakaki et al., 2008), our semantic annotation focuses on the relation between adjacent text spans. The text spans may contain single or composite discourse segments. The current scheme uses a dozen of labels under four classes: TEMPORAL (Sequence, Synchronous, Circumstance), CONTINGENCY (Reaction, Causal, Condition, Hypothetical, Fantasy), COMPARISON (Contrast, Concession), and EXPANSION (Elaboration, Justification).

The annotation is in progress and we are checking on the inter-annotator agreement. The resulting annotated corpus will provide a useful resource to back up our research on story analysis and story understanding from the linguistic, computational, and cognitive perspectives.

The Current Study

In the current study, we compare different versions of the same stories in Aesop's Fables for the variation in story structure, and more importantly, how such variation in structure interacts with the semantic relations among discourse segments so that the lesson of a story is preserved. We intend to study this structure-semantic interface in fables in terms of the surface linguistic features based on our annotated corpus.

We choose ten well-known stories in Aesop's Fables (e.g. the Hare and the Tortoise, the Fox and the Grapes, etc.), each with a short and a long version in both English and Chinese. The various versions of the same story are analysed for their surface linguistic features (e.g. lexical choices, rhetorical devices, use of dialogues, etc.), narrative structure and discourse relations, to see what multiple versions of a story might have in common structurally and how they vary linguistically otherwise, to allow storytellers to convey the intended message invariably by fleshing a basic skeleton of a story with different lexico-grammatical features, rhetorical devices, and discourse strategies.

In the presentation, we will first introduce the corpus and its annotation, and then present the results of our analysis.

Acknowledgements

The work described in this paper was supported by a grant from City University of Hong Kong (Project No. 7008064).

References

Grosz, B.J. and Sidner, C.L. 1986. Attention, Intention, and the Structure of Discourse. C*omputational Linguistics,* 12(3),175-204.

Hasan, R. 1996. The Nursery tale as a genre. In Cloran, C., Butt, D. and Williams, G. (Eds), *Ways of Saying: Ways of Meaning*. London: Cassell.

Mandler, J.M. and Johnson, N.S. 1977. Remembrance of things parsed: story structure and recall. *Cognitive Psychology,* 9, 111-151.

Mann, W.C. and Thompson, S.A. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*.ISI Research Report ISI/RS-87-190, University of Southern California.

Miltsakaki, E., Robaldo, L., Lee, A. and Joshi, A. 2008. Sense Annotation in the Penn Discourse Treebank. C*omputational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol.4919*, 275-286.

Rumelhart, D. 1975. Notes on a schema for stories. In Bobrow, D. and Collins, A. (Eds.), *Representation and Understanding: Studies in Cognitive Science*. New York: Academic Press.

Toolan, M. 2009. *Narrative Progression in the Short Story: A corpus stylistic approach*. John Benjamins Publishing Company.

Wilensky, R.1982. Points: A theory for the structure of stories in memory. In Lehnert, W.and Ringle, M. (Eds.), *Strategies for Natural Language Processing*. Hillsdale, NJ: Lawrence Erlbaum, 345-375.

EFL Students' Perceptions of Corpus-Tools as Writing Aids

Shu-Li Lai

Center for General Education
National Taipei College of Business, Taiwan

Keywords**:** corpus tools, EFL writing, parallel concordances, collocation behaviors, self-efficacy

Introduction

Recent studies have suggested the potential of corpus tools in vocabulary learning and EFL writing. However, little is known regarding how EFL writers perceive these tools as writing aids. To better understand the question, this study investigated 14 college students' perceptions of corpus tools right after their application of these tools to the writing tasks, in which four online corpus tools, including monolingual and bilingual concordancers and collocation retrieval systems were provided along with two online dictionaries. After two tool-training sections, students performed three timed-writing tasks online in three consecutive months and received individual recall interviews after each writing task. The interviews with the student writers served as the major source of data. By eliciting students' perceptions and comments right after they experienced these tools during the writing process, the results obtained revealed the writers' perceptions regarding how corpus tools contributed to their writing.

Results

Overall, the students' perceptions of the corpus tools as writing references were positive. First, students believed that the corpus tool was a complement to the bilingual dictionary in that a bilingual dictionary such as Yahoo Online Dictionary did not always provide enough sample sentences. Without sample sentences, students reported that they had difficulties learning the usage of the words. In contrast, corpus tools often provide sufficient example sentences. These sentences provide contextual clues which help them to select a word that is appropriate for the intended context; the sentences also serve as examples for the students to learn grammatical structures.

Second, students believed that the bilingual feature of the corpus tool was helpful and crucial. For one thing, it allowed bi-directional searches; for another, the Chinese parallel concordances provided clues to help the students quickly find the related information. In this study, the bilingual concordance was used much more often than the monolingual concordance. This indicated that a bilingual tool contributed more to the writing than a monolingual one. As commented by one student, whether the tools provided bilingual example sentences was his main concern when deciding the type of tool to use. According to this study, the bilingual information eased the consultation process and gave students with different language proficiency levels a chance to adopt the corpora good for them to perform the writing task.

Third, several students commented on the way the monolingual VLC Web Concordancer presented its results and found it to be a powerful feature. When VLC presents its results, the

keywords are aligned in the middle, highlighted in red. In addition, the words following the keyword are listed in alphabetical order, which not only made it easy for student writers to see the patterns but also drew their attention to the structures of which they had not been aware of before.

The Collocation Retrieval System was reported to be a helpful and user-friendly tool. Through several mouse clicks, the student writers could find the collocation behaviors of the target words and the frequency of each collocation. Students reported that the frequency of each collocate was helpful, especially when they could not decide which word to use after observing the search results.

The qualitative data also revealed how consulting corpus helped the students to increase their confidence in writing, particularly the wording. Having access to the corpus tools, most students believed that they could now use words in a more concise and accurate manner. As one student pointed out, corpus tools provided him with more information so that he could compare the findings with the result found in the dictionary. Consequently, his wording became more accurate.

Discussion

The results indicate that the availability of sufficient example sentences plays a crucial role in EFL writing. They provide contextual clues and display key words in context which help students figure out the meaning of the target word and differentiate words that are close in meanings. Additionally, these sentences provide syntactical information and act as models for usage of the keyword. Students imitate how native speakers structure sentences by either borrowing the sentence directly or making some minor modifications.

It is surprising that students with different proficiency levels all managed to use the corpus tools in some way, and most of them held a positive view of the tools. This contradicts the findings of Yoon and Hirvela's study (2004). Their questionnaire data showed that about 60% of the students from an intermediate and advanced class found observing the incomplete sentences in concordance output difficult; approximately 40% of the students found the corpus texts too difficult to read. In their study, students only had the access to a monolingual concordance, which may explain the differences between the results of their study and mine. In the present study, students were allowed to use both the monolingual and bilingual corpus tools. It is very likely that the bilingual feature of the corpus tools has made it easier for EFL writers to retrieve information from concordance lines.

Students' perceptions of the presentation of the corpus results by VLC require further discussion. The way VLC presents the information carries some theoretical and pedagogical implications. It helped students to "notice" patterns, and drew their attention to the collocation behaviors, grammatical patterns, and related meaning. According to Schmidt, "SLA is largely driven by what learners pay attention to and notice in target language input" (Schmidt, 2001, p.3). Integrating a corpus tools such as VLC into EFL training could not only solve students' writing problems but also enhance learning via its form of presentation.

Finally, students' successful corpus consultation experiences seem to have helped them to gain self-efficacy in writing. They believed that more corpus consultation would result in a higher accuracy rate. When one tool failed to provide them with satisfying information, they tried another; when the concordance lines were too difficult, they read parts of the sentences or used the

Chinese parallel corpus to help them comprehend the sentences. They overcame numerous obstacles through perseverant effort. As one of the participants said, good use of these tools would help him solve all of the writing problems. This revealed students' confidence in the tools and in their ability to use the tools to solve problems they would face in writing. In fact, several studies have shown that there is a relationship between self-efficacy and writing performance (Pajares, 2003). Perception may accurately predict students' motivation and even their future academic achievements.

Conclusion

In conclusion, this study examined the students' corpus consultation experiences through recalls and comments obtained directly after they integrated these tools into their writing tasks. Overall, the participants greatly valued these tools and believed that having access to the corpus tools helped them to improve their wording and increase their confidence in writing.

References

Pajares, F. 2003. Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, 19 (2), 139-158.

Schmidt, R. 2001. Attention. In P. Robinson (Ed.), *Cognition and second language instruction.* Cambridge: Cambridge University Press, 3-32.

Yoon, H., and Hirvela, A. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13, 257-283.

Corpus Linguistics – A Trend in Compiling ESP Documents for College and University Students in Vietnam

Thi Hoa Binh Lam

Faculty of English, University of Languages and International Studies
HaNoi National University, Vietnam

Key words: ESP, VNU, ESL, corpus-based, discipline

Teaching ESP in Hanoi National University

Teaching English for Specific Purposes (ESP) has long been a great consideration of many colleges and universities in Vietnam in order to meet more and more sophisticated demand in various disciplinary communities as well as in the social life. Hanoi National University is one of the greatest academies in Vietnam where more than 20 ESP documents have been being used in teaching and learning. Most of the documents were written by teachers of English here.

ESP teaching in VNU can generally be divided into three periods: before 1986, from 1986 to 2000, and after 2000.

Before 1986, teachers and professors in disciplines were in charge of teaching ESP for students. That is not the time for teaching language skills but for reading and translation. There was no language course book specially designed for students. However, the students learnt many of technical terms and how to read documents from books in their majors in a practical way.

During the period from 1986 to 2000, the compilation of an ESP document was merely based on one supposedly principal subject and a limited number of English handbooks and textbooks supplied by teachers of each discipline. Many reading passages and exercises in such documents, therefore, were repetitive and not very informative.

Since 2000, with the bloom of PCs and informatics technology such as the internet and cable TV, while preparing an ESP document, most teachers have made use of the various sources to enrich language activities. There are over 30 ESP documents which have been compiled since then, for instance, English for Students of Tourism (1998), English for Students of Law (2002) English for Students of Sociology (2003), English for Students of Literature (2006), English for Students of Linguistics (2007), and many others.

However, there has never been any comprehensive needs analysis on ESP for language teachers (and there will hardly be any in the near future) to rely on to gather proper knowledge for students from specific faculties. All decisions on curriculum, documents and duration for teaching are subjective, mainly based on teachers' experience and some suggestions from specialists and professors teaching in such disciplines. Apparently, it is impossible to define the appropriate amount of knowledge, contents and language skills in ESP documents. This is one of the major problems in ESP teaching and learning in Hanoi

National University in particular and in most colleges and universities in Vietnam in general.

For the past 10 years, along with the transfer of training modal from annual to credit in most colleges and universities in Vietnam, colleges in Hanoi National University (VNU) have undergone many changes in training content, time duration, credits for training, personnel, ESP training. Many ESP curriculums have to be trimmed to adjust to the changes. Consequently, teaching as well as learning ESP can hardly be as effective as it used to be.

In order to meet the VNU's requirements for education improvement and those made by disciplines in its colleges, designing appropriate ESP courses in conformity to the different levels of student, the amount of impartation, the language focus on grammar, lexicon, skills, etc., to teach and learn effectively within a time limit is pivotally important in the present time.

Corpus-based studies on English language teaching

According to Kennedy (1992: 335), "corpus linguistics has held potential relevance for the teaching of languages because responsible language teaching involves selecting what it is worth giving attention to". In fact, in teaching English as a second language, we endeavor to bring out a true-to-life language by teaching learners generalized language elements which are frequently used. However, the selection of proper materials has always been a problem. Many solutions have been thought of including Johns' (1990:91) suggestion of a bank that store teaching materials for ESP. However, such a bank does not seem to work (Kristen Gatehouse, 2001). ESP materials need editing and renewing accordingly to latest researches in specific fields, so selection of topic, source, vocabulary appropriateness, especially up-to-date scientific information takes even much more time.

Since the earliest and most significant research of 4.5-million-word corpus of Thorndike (1921) used as pedagogical materials for teaching reading, many other researches have been made, for instance, H.E. Palmer (1933), West (1953) studied on collocation, Coxhead (2002), Romer (2004) on vocabulary; George (1963b,c), Ota (1963), Dusková Urbanová (1967), Kránský (1969), Joos (1964), Coates (1983), Mint (2002) studied grammar; Kennedy (1978), Salager-Meyer (1990), Altenberg (1987), Fang (1990) with researches on syntax and semantics, Belz (2008) and Mollering (2001, 2004) studied on pragmatic and discourse features. The number of researches on corpus linguistics (CL) and their application in teaching and learning English as second language for decades indicates that corpus linguistics and corpus analysis have become more and more effective and indispensable approaches in language studies.

A corpus-based study on compiling ESP documents

In our research on corpus-based ESP compilation, we find the inter-influential elements in integrated course design (Fink, 2007) important and should be put into account. Our study is also based on the theories of curriculum design, instructions and assessment by Fink (2007), Grant Wiggins & Jay Mc Tighe (1998, 2005), especially Grant Wiggins and Jay Mc Tighe's "backward design" which seems to be appropriate to our University's status quo. The design does not require costly need analysis but defines the desired results in the first place. After designing the curriculum based on acceptable data, planning learning and teaching experience are considered. The process is suitable for designing English teaching

program of various levels in general and designing ESP programs effectively in particular.

Our corpus-based research aims to tailor ESP programs which fit the time allowance and correspond to the common curriculum based on the calculation of the frequency of words/phrases' occurrence in the corpus built from books, documents and information in a specific discipline. The focal registers will then be reviewed in contrast with another corpus of vocabulary and grammar in GE program (General English program which is being applied for the freshman and sophomore in VNU) to eliminate the most familiar which have been well comprehended. The core registers will then be used to decide the appropriateness of information sources, their content for an ESP document. Furthermore, it will be also helpful in selecting proper exercises and activities for listening, speaking, reading, writing and translation.

References

Belz, J. A. and Vyatkina, N. 2008. The *pedagogical mediation of a developmental learner corpus for classroom- based language instruction*, Gale Group.

Dang, T.T.T. 2007. English for Students of Linguistics, College of Social Sciences and Humanities.

George, H.V. 1963. "A verb-form frequency count", *English Language Teaching* 18(1), 31-37.

George, H.V. 1963. *A verb-form frequency count: application to course design,* Hyderabad: Central Institute of English, Monograph 2.

George, H.V. 1963. *Report on a verb-form frequency count,* Hyderabad: Central Institute of English, Monograph 1.

Howard, J. 2007. "Curriculum Development", Center for the Advancement of Teaching and Learnin, Elon University

Howard, J. 2007. Curriculum development. http://www.pdx.edu/sites/www.pdx.edu.cae/files/Howard.pdf

http://www.tlumaczenia-angielski.info/linguistics/corpus-linguistics.htm

Hutchinson, T. and Waters, A. 1987. English for Specific Purposes: A learning-centered approach. Cambridge: Cambridge University Press.

Kennedy, G. D. 1992. "Preferred ways of putting things with implications for

Kennedy, G.D. 1990. "Collocations: where grammar and vocabulary teaching meet". In: Anivan, S. (Ed.), *Language teaching methodology for the nineties.* Singapore: RELC Anthology Series, 215-229.

Kristen Gatehouse. 2001. Key Issues in English for Specific Purposes (ESP) Curriculum Development. *The Internet TESL Journal*, Vol. VII, No. 10, October 2001 http://iteslj.org/ ; http://iteslj.org/Articles/Gatehouse-ESP.html

Lam, Q.D. 2002. *English for Students of Law*, VNU Publishing House.

Lam, T.H.B. 2007. *EGEO 1: An advanced course for students of Geography* (Volume 1), University of Education Publishing House, Hanoi, Vietnam.

Lam, T.H.B. 2008. *ESP Curriculum: English for Students of Psychology*. College of Social Sciences and Humanities, Hanoi National University.

Lam, T.H.B. 2009. *EGEO 2 – An advanced course for students of Geography* (Volume 2), University of Education Publishing House. Hanoi, Vietnam.

Lam, T.H.B. 2009. *ESP Curriculum: English for Students of Social Work*. College of Social Sciences and Humanities, Hanoi National University.

Lam, T.H.B. 2009., *ESP Curriculum: English for Students of Sociology*. College of Social Sciences and Humanities, Hanoi National University.

language teaching", in: Jan Svartvik (ed.), *Trends in linguistics*, 335-373. Berlin: Mouton De Gruyter.

Leech, G. 1992. Corpora and theories of linguistic performance. In Svartvik, J. (Ed.), *Directions in corpus linguistics: proceedings of Nobel symposium 82,* Berlin and New York, Mouton de Gruyter, 125-148.

Mackay, R., and Palmer, J. (Eds.). 1981. *Languages for Specific Purposes: Program design and evaluation*. London: Newbury House.

McEnery, T. and Wilson, A. 2001. *Corpus linguistics*, Edinburg University Press.

McEnery, T., Xiao, R. and Tono, Y. 2006. *Corpus-based language studies: an advanced resource book*. Routledge Taylor & Francis Group

McEnery,T. and Wilson, A. 2001. Corpus linguistics, Edinburgh University Press.

Nguyen, H.L and Lam, T.H.B. 2003. *English for Students of Sociology*, College of Social Sciences and Humanities, Hanoi National University.

Nguyen, H.L and Luong, H.N. 2007. English for Students of History, College of Social Sciences and Humanities, Hanoi National University.

Nguyen, H.L. et al. 1998. English for Students of Tourism, Hanoi National University, Hanoi National University.

Nguyen, V.T., and Lam, T.H.B. 2003. About a new ESP syllabus for students of Geography. In *Journal of Science* (Vol 3, 98-101). Hanoi University of Education.

Ota, A. 1963. *Tense and aspect of present-day American English*. Tokyo: Kenkyusha.

Palmer, H. E. 1933. *Second interim report on English collocations.* Tokyo: Institute for Research in English Teaching

Pham, T.T. 2007. English for Students of Literature, College of Social Sciences and Humanities.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language,* London: Longman.

Svartvik, J. (Ed.) 1990. *The London-Lund Corpus of Spoken English: Descirption and research,* Lund: Lund University Press.

Tu, N.A. 2007. English for Students of Philosophy, College of Social Sciences and Humanities.

Tyler, R. W. 1949. *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Wiggins, G. 2002. Toward assessment worthy of the liberal arts: The truth may make you free, but the test may keep you imprisoned. Mathematical Association of America. Available at http://www.maa.org/SAUM/articles/wiggins_appendix.html Wildman, T. M. (Spring 2007).

Yogman, J., and Kaylani, C. 1996. ESP program design for mixed level students. *English for Specific Purposes,* 15, 311-24.

Building and Using English-Vietnamese Parallel Corpora (EVPC)

Van Le

Inverness Technologies – Canada

Keywords: English-Vietnamese parallel corpora, corpus linguistics, dictionary

Introduction

In the last twenty years, a significant number of research projects have focused on Corpus Linguistics. We can find many applications of Corpus Linguistics: in teaching and learning English, in translation, in linguistics and in lexicography. Corpus Linguistics applications are found in many languages, particularly European languages. In the last decade there were almost no corpus linguistics articles available in Vietnamese. However, we have recently been able to see some interesting reports. (Kilgarriff et al on Corpus Factory, and Phuong Le-Hong on Vietnamese POS tagger, for example).

This is just the tip of the iceberg, compared to the large amount of corpus linguistic information available in other languages. We have to do more to exploit the large number of resources available on the Internet, with most of these being in English. In order to learn more about Corpus Linguistics, Vietnamese readers must learn English. Therefore we need to create many EV translation memories, and EV Parallel Corpora to help readers to speed up their learning. Parallel corpora have been a topic of research for many groups in Vietnam. The results of this research have been published on the Internet. However, the contents of these parallel corpora only reflect part of the language. For example, the Sealang project exploits the contents of simple dictionaries published in Vietnamese, while Dinh Dien focuses on science and conventional fields. Other articles report on the process of building an annotated corpus, including Thi-Minh-Huyen Nguyen and allies on Building a Large Syntactically-Annotated Corpus of Vietnamese, and Dang & Ho on the process of automatic retrieving information from the Internet for building an EV parallel corpora.

In this article we present the principle of building one-million lines of English-Vietnamese Parallel Corpora (EVPC). Part of this EVPC has been tested successfully on the server of the Sketch Engine. The content of this EVPC covers most topics associated with daily activities.

Part One: Building the EVPC

There are four steps in the process of building the EVPC: Locating the resources, Pre-Processing (file and font converting), Alignment, and Post-Processing

Part Two: Content of the EVPC

As mentioned above, parallel corpora have been the topic of research by many groups in Vietnam. The results of this research have been published on the Internet. However, the contents of these parallel corpora only reflect part of the language. As noted, the Sealang project exploits the contents of simple dictionaries published in Vietnamese, while other projects have focused on Information Technology.

In building our EVPC, we intend to cover most language needed: from kindergarten to advanced research, science and technology to law and legal subjects, and local to international activities.

The main topic areas for this EVPC are: Education and Training, Law and Order, Medical and Health, Business and Commerce, Individual and Social Rights and Responsibilities, Science and Technology, Human Resources, USA Life, Vietnamese Culture, International Relationships

Part Three: Using the EVPC

For Linguistics:The EVPC has been tested with the Sketch Engine. The Sketch Engine is a corpus processing system developed in 2002 (Kilgarriff and Tugwell, 2002; Kilgarriff et al., 2004). The main components of the Sketch Engine are concordances, word sketches, grammatical relations, and a distributional thesaurus.

For Translation:

We have tested this EVPC on different Computer Aided Translation tools (SDLX, Trados). The EVPC can be converted into suitable formats for use on electronic.

For Language Study:As mentioned in Part Two, we have American English texts as well as British texts. This provides us the resources for teaching and learning about all aspects of English language. With many different topics, EVPC can be used as a resource for English for specific purposes. For Vietnamese, we do have grammar rules, but we can find different styles in articles written by those living inside and outside the country, and in articles from the North and from the South of Vietnam. EVPC can be used to deal with problems that occur in Vietnamese students' writing that occurs due to the influence of their native language. We also use the word list within the English section as a group of seed words for creating larger EV Parallel Corpora by using special, online dictionary creating software as well as Google Translate. With minor editing, this expanded version can be used as an excellent resource for learning and teaching.

Contrastive Linguistics:

The term 'contrastive linguistics' was suggested by Whorf (1941) and was defined as "a sub discipline of linguistics concerned with the comparison of two or more languages or subsystems of languages in order to determine both the differences and similarities between them" (Fisiak, 1981:1). The EV Parallel Corpora is a valuable source for Contrastive Linguistics. Contrastive Linguistics focuses on the translation product and tries to discern similarities and differences between English and Vietnamese by comparing parallel texts. English for Specific Purposes:As mentioned earlier, within each topic area of the EVPC, there are many sub-sections. By using Corpus Linguistics analysis tools, we can also identify specific word lists. This establishes an excellent resource for creating English Glossaries and Dictionaries for specific purposes. Word lists can be used for creating keyword lists to show students differences in vocabulary, style, and other components of language. EVPC can also be used to teach English collocations and N-Gram. One of the Corpus Linguistics analysis tools that is helpful to discover the behaviour of words is a concordancer. It finds all the occurrences of search words in a corpus and presents them in the form of a list called a concordance.

Conclusion

The English-Vietnamese Parallel Corpora was started in 2000. Over the years we have not only increased the number of lines but also created many tools for exploiting the parallel corpora. Following is the list of tools and utilities which have been created: Multilanguage dictionary, TransMem, VinaCorpora, LQV.exe

Acknowledgements

We would like to thank Dr. Bao Phi, for creating transMem, VinaCorpora and LQV.exe. We also would like to thank Mrs. Lesley Chow and Mr. Paul Gowan for reviewing the earlier version of this article.

References

KilgarriffAdam, Siva Reddy, Jan Pomikálek, Avinesh PVS. A Corpus Factory for many languages. http://www.kilgarriff.co.uk/publications.htm.

Phuong Le-Hong. Vietnamese POS tagger. http://www.loria.fr/~lehong/publications.php.

Sealang project. Vietnamese Corpus and Dictionary. http://www.sealang.net/archives/mks/pdf/35:21.pdf.

Dinh Dien. Building an annotated English-Vietnamese parallel corpus. http://pseal.org/vietnamese/corpus.htm.

Thi-Minh-Huyen Nguyen and allies. Building a Large Syntactically-Annotated Corpus of Vietnamese. http://www.aclweb.org/anthology-new/W/W09/W09-3035.pdf.

Dang, V.B. and Ho, B. 2007. Innovation and Vision for the Future, 2007 IEEE International Conference,.

Corpora and Data-Driven Learning (DDL) Approach

Dong Ju Lee

Division of National English Ability Test
Korea Institute for Curriculum and Evaluation, Korea

Keywords**:** corpora, DDL approach, classroom concordancing, corpus-based materials, computer-aided error analysis(CEA)

Introduction

This study provides a brief introduction to corpora and the Data-Driven Learning (DDL) approach by introducing key concepts and major principles. It also presents two types of DDL – soft and hard versions – and exemplifies various types of activities which have been commonly used in the soft version of DDL materials. Then, the final part describes students' and teachers' attitudes towards the DDL approach to learning and teaching writing and grammar in the Korean EFL context.

Key concepts and major principles of DDL

According to Johns and King, DDL, classroom concordancing, is defined as:

> The use in the classroom of computer-generated comcordances to get students to explore the regularities of patterning in the target language, and the development of activities ans exercises based on concordance output (Johns and King, 1991: iii).

In a DDL classroom, the language learners are generally provided with concordance data as 'enhanced language input' (Kettemann, 1997) to enrich theor 'language awareness' (Van Lier, 1995) and/or to lead to 'consciousness-raising' (Ellis, 1992; Rutherford, 1987; Sharwood-Smith, 1990). As a way of providing 'form-focused instruction' (Granger and Trible, 1998), classroom concordancing helps the learners generalize patterns and use of target language items by themselves.

The most oticeable principles of DDL approach can be summarised as follows: (1) The language input in DDL classrooms is presented in the form of concordance lines which are authentic language samples taken from a pedagogically useful corpus; (2) DDL is a new way of discovery learning: the learners are strongly encourgaed to work on concordance data to find out language patterns and use by themselves; (3) DDL is a strong form of grammatical consciousness-raising which can be particularly useful in raising learners' attention to particular language features and developing their inductive learning strategies as a language learning tool; and (4) The learners are at the centre of the classroom while the teacher takes the role of a coordinator or facilitator of student-centred research rather than an authoritative language input provider.

Two types of DDL

Leech(1997: 10) says corpora can be used in the form of a concordance in language teaching in two ways: the 'soft version' involves teacher-designed concordance materials in the form of printouts or worksheets; while the 'hard version' involves learners conducting autonomous or

independent concordancing themselves by directly accessing a concordancing program using computers, CDs or web-based online tools.

In the soft version, the teacherusually has access to a corpus and the relevant software, prints out concordance samples from the corpus, and designs tasks and activities. The, learners are encouraged to work with those teacher-designed corpus-based materials (see Bernardini, 2004; Granger and Tribble, 1998; Tribble and Jones, 1990). In the hard version, on the other hand, learners are required to have direct access to computer and corpus facilities and have the skills to use them. In this version, activities and tasks can be devised in different ways: they can be created by the teacher, incorporated into a CALL program, or selected by the learners, with or without the instructor's management.

Students' and teachers' attitudes towards DDL approach

Based on the computer-aided error analysis (CEA) of a Korean learner corpus of English(Lee, 2008), teacher-designed corpus-based materials were developed as DDL teaching lessons to remedy the learners' most frequent error types and tried out on Korean secondary school English classes. Qualitative instruments (classroom observations and in-depth interviews), as well as a questionnaire were employed to investigate the students' and teachers' attitudes to the DDL approach. The results encouragingly revealed that both the students and teachers felt relatively positive about this type of pedagogical approach. The data also shed light on problems with the design of DDL activities and materials, pointing to improvements which could be made in future versions.

References

Bernardini, S. 2004. Corpora in the classroom: An overview and some reflections on future developments. In Sinclair, J. (Ed.), *How to use corpora in language teaching*. Amsterdam: John Benjamins, 15-36.

Ellis, R. 1992. Grammar teaching – Practice or consciousness-raising? In Ellis, R. (Ed.), *Second language acquisition and second language pedagogy*. Clevedon: Multilingual Matters, 232-241.

Granger, S and Tribble, C. 1998. Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In Granger, S. (Ed.), *Learner English on computer*. London: Longman, 199-209.

Johns, T and King, P. (Eds.). 1991. *English Language Research Journal Vol.4: Classroom Concordancing*. Birmingham: The University of Birmingham.

Kettemann, B. 1997. Concordancing as input enhancement in ELT. In Lewandonska-Tomaszczyk,B. and Melia, P. J. (Eds.), *Proceedings of the 1st International Conference on Practical Applications of Language Corpora (PALC)*. Łódź, Poland: Łódź University Press, 63-73.

Lee, D. J. 2008. A computer-aided error analysis (CEA) of a Korean learner corpus. *Modern English Education*, 9 (1), 73-89.

Leech, G. 1997. Teaching and language corpora: A convergence. In Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (Eds.), *Teaching and language corpora*. New York: Addison Wesley Longman, 1-23.

Rutherford, W. 1987. *Second language grammar: learning and teaching*. London: Longman.

Sharwood-Smith, M. 1990. Consciousness-raising and second language learner. *Applied Linguistics*, 11 (2), 159-168.

Tribble, C and Jones, G. 1990. *Concordances in the classroom: A resource book for teacher*. London: Longman.

Van Lier, L. 1995. *Introducing language awareness*. London: Penguin English.

L2 Acquisition of Adverb *Cai* in Mandarin Chinese

Jen-I Li[1], Miao-Ling Hsieh[1,2], and Hao-Jan Chen[1,2]

[1]*Department of English*
[2]*Mandarin Training Center*
National Taiwan Normal University, Taipei, Taiwan

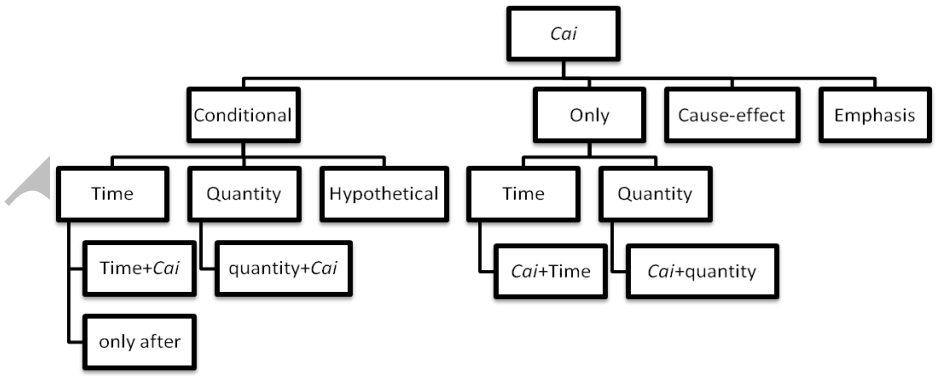
Keywords: interlanguage, learner corpus, online writing tests, focus particle, Mandarin Chinese, CSL

Purpose of the Study

The purpose of this study is to examine L2 adult learners' acquisition of the adverb *cai* in Mandarin Chinese using the data from a corpus consisting of online writing tests of different genres conducted in the Mandarin Training Center of National Taiwan Normal University.

Analysis Framework

Cai is an adverb, functioning as a focus particle. It has two basic meanings—'just now' and 'only then' (Li and Thompson, 1981: 332). However, later research shows that the meanings and functions developed from these two basic meanings are quite complicated (Biq, 1984, Lai, 1995, Wang, 2003, etc.). A recent study of L1 acquisition of *cai* in Mandarin Chinese identifies four domains for the use of *cai*: time, quantity, condition and restriction ('only') (Yang, 2009). Based on the production data by L2 learners, we categorize the occurrences of *cai* into four major types with subcategories by taking into consideration not only *cai*'s semantic and pragmatic functions, but also the syntactic complexity of the *cai* sentences. The four major types are conditional, 'only', cause-effect, and emphasis. Conditional *cai* is further divided into time, quantity and hypothetical, while 'only' *cai* is subcategorized into time and quantity. All the types are exemplified below.



Data Analysis and Results

The corpus used in this study contains 1 million characters. They come from the online writing tests taken by L2 adult learners in the Mandarin Training Center at National Taiwan Normal University. The tests are given at four levels according to the *Common European Framework of Reference* (CEFR) —with A2 the lowest and C1 the highest level, B1 and B2 the levels in between. All the learner language produced for each level is collected and analyzed. In the corpus, 849 instances of *cai* are annotated as adverb. 51 of them are not interpretable. The 799 instances left are analyzed according to the types mentioned above.

Our first finding is that the conditional *cai* (83.6%) was used most frequently, followed by cause-effect (11.3%), 'only' (4.4%) and emphasis *cai* (0.8%) respectively, as illustrated in Table 1. With respect to the error rate, the lowest was found in emphasis *cai* (0.0%), and the

highest in 'only' *cai* (14.2%), with the other two in between (cause-effect 13.3%, conditional 11.8%), as shown in Table 2. When the learners' proficiency levels are taken into consideration, some developmental tendencies are revealed, as shown in Table 3. First, the error rate became lower when the proficiency level was higher. Second, within the conditional type, the percentage of occurrence of time *cai* became lower and that of the hypothetical became higher when the proficiency level was higher.

In the paper, the results mentioned above will be discussed in detail and compared with the results of L1 acquisition found by Yang (2009). The errors made by the learners will also be analyzed and discussed in order to find out why such errors occur. Through this study, we hope that we can have a better understanding of the meanings and functions of the adverb *cai* and have some contribution to teaching/learning Mandarin Chinese as a second/foreign language.

Table 1 Total occurrences of the four types of *cai*

	Conditional	'Only'	Cause-effect	Emphasis
Total	668 (83.6%)	35 (4.4%)	90 (11.3%)	6 (0.8%)

Table 2 Error rates of the four types of *cai*

	Conditional			'Only'		Cause-effect	Emphasis
	Time	Hypothetical	Quantity	Time	Quantity		
Correct	339	234	16	20	10	78	6
Incorrect	37	42	0	4	1	12	0
Total	668			35		90	6

Table 3 Occurrences of the four types of *cai* in different proficiency levels

Level	types / accuracy	Conditional			'Only'		Cause-effect	Emphasis	Total
		Time	Hypothe-tical	Quantity	Time	Quantity			
A2	Correct	102	9	2	5	0	0	0	118(88.8%)
	Incorrect	7	8	0	0	0	0	0	15(11.2%)
	Total	109(82%)	17(12.8%)	2(1.5%)	5(3.8%)	0(0%)	0(0%)	0(0%)	133(100%)
B1	Correct	148	99	4	11	8	41	1	312(84.8%)
	Incorrect	21	24	0	4	1	6	0	56(15.2%)
	Total	169(45.9%)	108(33.4%)	4(1.1%)	15(3.8%)	9(2.2%)	47(12.8%)	1(0.3%)	368(100%)
B2	Correct	77	99	8	3	2	32	5	225(89.3%)
	Incorrect	9	9	0	0	0	6	0	24(10.7%)
	Total	86(34.5%)	108(43.4%)	9(3.2%)	3(1.2%)	2(0.8%)	37(14.9%)	5(2.0%)	249(100%)
C1	Correct	12	27	2	1	0	6	0	48(98%)
	Incorrect	0	1	0	0	0	0	0	1(2%)
	Total	12(24.5%)	28(57.1%)	2(4.1%)	1(2.0%)	0(0%)	6(12.2%)	0(0%)	49(100%)

Acknowledgements

This project is supported by 'Aim for Top University Plan' sponsored by the Ministry of Education, Taiwan, R.O.C.

References

Biq, Y.-O. 1984. The Semantics and Pragmatics of *Cai* and *Jiu* in Mandarin Chinese. Ph.D. Dissertation, Cornell University.

Lai, Huei-ling. 1995. *Rejected Expectations: the Scalar Particles CAI and JIU in Mandarin Chinese.* Ph.D. dissertation. University of Texas at Austin.

Li, C. N. and S. A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Wang, N. 2003. Syntactic-semantic analysis of "cai" sentences. *Journal of Chinese Information Processing,* 17(1), 38-45.

Yang, X.L. 2009. *Focus and Scales: an Experimental Study of L1 Acquisition of Cai and Jiu in Mandarin Chinese.* Beijing: Beijing University Press.

English Measure Noun Phrases: A Usage-based Approach

Jooyoung Lim and Jong-Bok Kim

Kyung Hee University, Seoul

Keywords: measure noun phrases (MP), selectional restriction, concord, modification, Korean EFL learners, authentic usage

Measure noun phrases (MP) canonically realized in the form of X of Y exhibit grammatical complexities, causing high level of learning difficulties to most of the EFL learners. MPs individuate and give classificatory information about Y, but induce grammatical indeterminacy with respect to number concord, selectional restriction and modification patterns. For example, subject-verb agreement is determined either X or Y as in the attested corpus examples *A group of girls are sitting on a porch trying to sing* vs. *A bunch of grapes hangs too high*. The selectional restrictions of the verb can be also satisfied either of the two as in *I can't hold a glass of water* vs. *I was sipping a glass of water*. We can hold a glass (X) and sip water (Y).

As an attempt to better understand such complex grammar of MPs, we have performed a corpus search. In so doing, we have first classified MPs into five groups (container-measures (e.g. *glass*), standard-measures (e.g. *gallon*), dimensional-boundaries (e.g. *ribbon*), configuration type (e.g. *stack*) and collection-of-members (e.g. *group*), and then performed a in-depth corpus search using the COCA (Corpus of Contemporary American English) with about 410 million words of spoken as well as written texts. At the initial stage of search, we extracted 20 measure words defined in *Longman Grammar of Spoken and Written English* (Biber and Harlow, 1999) and identified the frequency of each one, totaling 33,241 tokens. Of these, in terms of the group frequency, the collection-of-members type yields the highest frequency (53.98%) followed by dimensional (21.55%), standard-measure (10.74 %), configuration (6.30 %), and container-measures (7.42%). The top 4 frequent words we identified in the collection-of-member are *group, set, series, bunch.* In addition, the most frequent words in the four remaining groups are *cup, foot, piece, pile,* respectively. Of these 33,241 instances, we have extracted 100 examples for each measure word and checked how they behave with respect to the three main properties – concord, selectional restriction, and modification.

Our corpus data also prove the complexity of MPs, and also tell us the importance of understanding the grammar in the spirit of grammatical interfaces. In terms of number concord, the verb or pronoun agrees with X not with Y, indicating that X canonically serves as the agreement head. However, collection-of-members type allows the concord with Y, due to the semantic properties of Y. X in general determines the subject-verb agreement, but many instances of the collection-of-members type allows Y to function as the agreement head. We attribute this to the fact that the subject-agreement relies on the construal of the

subject NP, not just its morpo-syntactic feature. In terms of selectional restriction, container-measure types make X as the head unlike the other four types. As for the selectional restriction, Y is the main semantic head. This is due to the core semantic features are in Y rather than in X. In terms of the modification, adjectives can modify both X and Y. Modification for X is more frequent than Y. The proximity condition for modification seems to prefer this.

As a way of checking what kind of grammatical features for MPs Korean EFL learners encounter, we have also checked five high school textbooks currently in use but found out the learners are surprisingly given few opportunities to learn the nature of MPs. The study implies that proper language understanding as well as learning cannot rely on just one aspect of the grammar (e.g., syntax, semantics, or pragmatics), but rather requires integrative and active methods that place strong emphasis on authentic usages.

References

Biber, D. and Harlow, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Huddleston, R. and Pullum, G. K. 2002. *The Cambridge Grammar of the English Language.* Cambridge University : Cambridge.

Jackendoff, R. S. 1986. Quantifiers in English. *Foundations of Language* 4: 422-442

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longman

Stickney, H. 2007. From Pseudopartitive to Partitive. In Belikova, A. et al., (Eds.), *Proceeding of the 2nd Conference on Generative Approach to Language Acquisition North America*. 406-415.

On Chinese University Students' English Metadiscoursal Oral Chunks

Weiyan Lin

South China Normal University, Jinan University, China

Keywords: metadiscourse, machine-cut oral chunk, metadiscoursal oral chunk, oral corpus

In light of the phraseological tendency of language, this dissertation sets out to investigate features of Chinese university students' metadiscoursal oral chunks (CH MOCs). Two key notions are highlighted in this study: metadiscourse and machine-cut oral chunks (OCs). Metadiscourse is a central pragmatic construct referring to non-propositional elements in a discourse which signal discourse organization and writer /speaker stance towards the propositional message. Metadiscourse is ubiquitous in English and yet largely neglected in EFL teaching, the result of which is students' neglect of its use in both their speech and writing, thus decreasing coherence and interactiveness of their English expressions. A number of studies have been conducted in recent years on metadiscourse, but not many have been on oral English, and even fewer have focused on machine-cut OCs. The latter refers to continuous recurrent sequences of words automatically retrieved from oral corpora. Previous studies have revealed some pragmatic features of machine-cut OCs in general, yet their metadiscoursal features need further exploration in aspects of theoretical justification, linguistic patterns and association between metadiscoursal features and OC frequency cut-offs. Moreover, in-depth investigations need to be carried out on metadiscoursal features of EFL learners' machine-cut OCs. The overarching goal of this research is to interpret Chinese university students' English metadiscoursal performance, especially in the aspect of utterance initiation, through an analysis of their machine-cut OCs in the hope of providing implications for university oral English teaching in China.

To this end, this study focuses on utterance-initiating MOCs based on target learner corpora (*i.e.,* LINDSEI-CH and SECCL 2.0, corpora of Chinese university students' oral English, and LINDSEI-10, an oral English corpus of university students' from 10 countries) and native-speaker reference corpora (*i.e.,* LOCNEC and ICE-GB-Spoken). Typical linguistic patterns of English native-speaker MOCs were first constructed by implementing the working procedure of constructing meaningful units proposed by Lexical Grammar in corpus linguistics. They then served as reference to reveal CH MOC features in terms of linguistic patterns, functions and association between MOC proportions and OC frequency cut-offs. Finally some distinctive CH MOC features was verified in larger EFL databases.

The study results are as follows. First, native-speaker OCs (NS OCs) were characterized with two typical linguistic patterns (*i.e.,* "conjunction/particle + subject" and "pronoun + epistemicity /attitude /degree /communication") and associated positively with MOC

proportions. Second, CH MOCs were underused in both types and tokens, formally with less diversified elements and marked by more dysfluencies, and functionally signaled less discourse-framing feature and carried a more assertive tone. In addition, a positive association was not found between CH MOC proportions and their OC frequency cut-offs. Verification of the last feature in larger EFL databases revealed that it was shared by OCs of larger-grouped Chinese university students and not by OCs of university students from ten countries.

Theoretically, this study revealed the tendency of high frequency machine-cut OCs to form units of metadiscoursal meanings, thus providing new empirical evidence for the phraseological tendency of language and enriching analyses on multi-types of phraseological units in corpus linguistic perspective. Methodologically, this study proved that by implementing the working procedure of Lexical Grammar in constructing meaningful units, particularly enlightened by its concepts of frequency-driven approach, colligation and semantic preference, linguistic patterns and functions of MOCs could be identified from high frequency machine-cut OCs within four-word chunks. Pedagogically, linguistic patterns of MOCs constructed and high frequency MOCs identified in this study can provide reference for the teaching of OCs and evaluation of EFL learners' metadiscoursal performance.

The Construction of Interlanguage Corpus of CSL Learners

Ya-Hui Lin[1], Feng-Yi Chen[1], Miao-Ling Hsieh[1,2] and Shu-hua Fang[1]

[1]*Mandarin Training Center,*
National Taiwan Normal University, Taipei, Taiwan

[2]*Department of English,*
National Taiwan Normal University, Taipei, Taiwan

Key words：interlanguage corpus, learner corpus, sentence-based corpus, CSL, error code, metadata

Introduction

A difficulty interlanguage studies face is the circumlocution and avoidance strategies that learners use. For example, very few resultative *ba* sentences in Mandarin Chinese, if any, are found in learner corpora. To solve this problem, the Mandarin Training Center (MTC) of National Taiwan Normal University launched a sentence-based learner corpus project in August 2010. This project aimed at collecting the data from MTC achievement tests (including A1 to B2 levels[7]) and doing related research. Despite the fact that the collected sentences based on the tests are not natural daily output, they are CSL learners' systematic output on demand. It is hoped that the compilation and analyses of such data will benefit the teaching and learning Chinese as a second language.

Data Collection

The achievement tests are administered at the MTC as screening exams for some courses at end of each three-month term to determine whether students move up to a higher level, stay in the current level or move half way back to the original level. The achievement tests that were keyed were those held from Aug. 2010 to July 2011. Only those answers in sentence form were included. With each file representing a student, now 4,722 files have been generated. Each file contains about 4-6 sentences. There are in total about 23,610 sentences (451,242 Chinese characters). Wrong characters written by the students are marked with various codes in the corpus.

A Research Example

From A1 to B2, 78 sentence structures are tested in the achievement tests. To make sure that those sentence structures are used, students are prompted to answer questions, re-organize words or rephrase the sentences, using the designated forms. For example, for the [*ba*…V-*de*]

---

[7] The major teaching materials at the MTC are mapped into the six major levels based on the Common European Framework of Reference (CEFR). Please refer to Hsieh and Lin (2008): The Implementation of the CEFR in Chinese Teaching. Paper presented at ACTFL, Nov. 21-23, Orlando, FL.

pattern (B1 level), the students were asked to use the pattern to rephrase the sentence in (1):

(1) Wo jia gebi shi KTV dian,chao-debudeliao, haiwo mei tian dou
I home neighbor be KTV store disturb:with:noise-DE a:lotcause I every day all
shui bu hao.
sleep not good
'A KTV store is next to my home. It is extremely noisy and makes me sleepless.'

A typical answer in (2) has an affectee introduced by *ba* and a result preceded by V-*de*. Crucially, the verb has to be highly transitive to be used in a *ba* construction. Among the three verbs, *hai* 'cause', *chao* 'disturb with noise' and *shui* 'sleep', only *chao* is a possible candidate.

(2) Wo jia gebi shi KTV dian, meitian ba wo chao-deshui bu hao.
I home neighor be KTV store every day BA Idisturb:with:noise-DE sleep not good
'A KTV store is next to my home. It is extremely noisy and makes me sleepless.'

The analysis of our corpus data shows that only 35% (51 out 145) of the students were correct in constructing a *ba* sentence, as shown in Table 1. Table 2 indicates that about 49.4% (42 out of 85) of the students fall in the V(erb) Choice category, which means they were wrong in choosing the wrong verb. If we count those who also made other mistakes, 72.9% (62 out of 85) of them were mistaken in their use of the verb. The results show that verb choice is very difficult for the students.

Table 1: The Ba Construction

Correct		No answer		Errors		Total	
51	35%	9	6.2%	85	58.6%	145	100%

Table 2: Error Types

O(bject)	V(erb) choice	R(esult)	W(ord) O(rder)	No *ba*	O, R	V, O	V, O, R	V, R	V, TW	V, TW, WO	V, WO
3	42	8	4	1	5	9	1	2	4	1	3

More discussion of the *ba* construction and more other examples will be given in the paper.

Acknowledgements

This project was sponsored by the National Science Council, Taiwan, R.O.C. (CNS99-2631-S-003011) from August 2010 to July 2011 and was later integrated into 'Aim for Top University Plan', supported by the Ministry of Education, Taiwan, R.O.C.

Abu Ghraib, Guantánamo, and the New York Times:

A Corpus-assisted Critical Discourse Analysis

Will Lingle

Department of English
Busan University of Foreign Studies, Busan, South Korea

Keywords**:** critical discourse analysis, corpus assisted discourse studies, systemic functional grammar, war on terror, torture

Introduction

The killing of Osama bin Laden has revived the debate over the use of torture in interrogating suspected terrorists, and produced a new debate over the descriptions of violent interrogations in the news media. This debate has spilled over onto the front pages. Responding to reader comments that questioned the absence of the term 'torture' in stories depicting US abusive interrogation techniques, The New York Times openly admitted that they have avoided using the word ‘torture’ in covering the Abu Ghraib and Guantánamo scandals.

This was done, they claimed, in order to avoid taking sides. While editorial policy allowed use of the word 'torture' in opinion pieces, it was not to be used in hard news stories. The policy was crafted in the interest of maintaining journalistic neutrality on a contentious issue. This paper presents a test of that claim of neutrality.

Methods

A corpus of New York Times front page stories on the torture scandal was compiled using the terms 'Abu Ghraib torture' and 'Guantánamo torture'. These terms were chosen to search for any uses at all where the term appeared in connection with the story of these two places where torture of US prisoners has been documented (Greenberg and Dratel, 2005, Jaffer and Singh, 2007).

The corpus was then analyzed using AntConc (Anthony, 2011) to profile the newspaper’s descriptions of US violence. Concordances, collocations, and clusters of all instances of the term 'torture@' were studied to determine if any factual descriptions existed where the term was used to describe US actions. Using categories from the Systemic Functional Grammar of Halliday and Matthiessen (2004), associated Actors (subjects of torture) were compared with identifiable Goals (objects of torture).

A transitivity analysis was also performed on a sample New York Times news text to identify Process types and their associated Participants: essentially, who and what were described in the text and what actions or states were described? The analysis focused

especially on Material Processes (verbs that describe physical actions) and their associated Actors (subjects) and Goals (objects) that described violent physical actions.

The results from this were then compared with other texts describing the physical treatment of US prisoners: a declassified government report on the abusive interrogation of a detainee, a report from Human Rights Watch where a US soldier described actions he witnessed and committed in handling prisoners, a report from Physicians for Human Rights on the treatment described by former prisoners, and declassified statements from Abu Ghraib prisoners describing detainee abuse they themselves witnessed. Finally, concordances, collocations, and clusters of Material Processes describing violence in these texts were in the corpus of New York Times texts.

Results

As the paper described, the New York Times did not use the term to refer factually to US actions, though it did use the term this way to describe the actions of other (enemy) states. The term is used uncritically when Saddam-era Iraqi guards are the torturers, but in any case where US actors are involved, the term appears in quotations from outside sources. Typically, these are quotations from Bush administration officials denying the use of torture by American forces.

Material Processes describing violence such as 'hit', 'beat', and 'strip' appear in a similarly clear pattern: Americans hit prisoners only in statements from others, never in factual descriptions by New York Times writers. This is not the case for non-American actors, however, whose violence is described uncritically.

Also different was the level of violence described: the human rights reports, prisoner statements, and government report all describe a similar level of brutal violence committed by US soldiers against their prisoners. New York Times reports, however, present a much milder picture. Violence is not often described, and when it is, it is described in very mild terms, and attributed to outside sources, frequently in association with official denials.

Conclusion

Overall, the New York Times did not present a neutral picture of US violence, given the level of factual documentation available to journalists during the period under study, 2005-2011. New York Times writers truly went out of their way to present a milder, more doubtful account of US torture of prisoners (a war crime under US and international law) than that required by the facts available. Far from being a he said, she said situation, factual evidence of US torture of prisoners has been accumulating, a fact the New York Times has reported, but still avoided using in its own news reporting. New York Times coverage of the torture scandal in fact represented a Bush administration view. This creates a very different picture of the abuse of detainees than that presented elsewhere, on a morally, politically and legally significant issue.

References

Anthony, L. 2011. AntConc (version 3.2.2.1) [computer software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Greenberg, K. and Dratel, J. (Eds.) 2005. *The Torture Papers: The Road to Abu Ghraib*. New

York: Cambridge University Press.

Halliday, M.A.K. and Matthiessen, C. 2004. *An Introduction to Functional Grammar* (3rd ed.) London: Hodder Arnold.

Jaffer, J. and Singh, A. 2007. *Administration of Torture*. New York: Columbia University Press.

Linguistic Devices to Elicit Elaborated Talk in Interview Dialogues:

The Case of *the Corpus of Spontaneous Japanese*

Masanobu Masuda[1]

[1]*Institute of General Education*
Koshien University, Japan

Keywords**:** Japanese interview dialogue corpus, conversation analysis, multi-unit questions, trouble-anticipatory questions, news receipts

Introduction

This paper analyzes a Japanese interview dialogue corpus with conversation analytic methods, and presents descriptions of the linguistic devices utilized in the dialogues to elicit elaborated talk: elaborate multi-unit questions, trouble-anticipatory questions and exaggerated news receipts. The description illustrates the participants' orientation in the corpus through sequential analysis.

Corpus

The corpus analyzed is *the Corpus of Spontaneous Japanese (CSJ)*, a large spoken language database for the development of the speech recognition system and the linguistic and phonetic study of spontaneous speech. (Maekawa, 2002)

The audio-recording of 32 two-party interview dialogues in the CSJ were analyzed. Because they have to continue talking for more than 10 minutes, the participants, especially the interviewers (IRs), are oriented to eliciting longer and elaborated talk. Three devices are utilized for that purpose, which is discussed below.

Elaborate Multi-unit Questions

The IRs often ask multi-unit questions, which do not just convey questions but involves complex communicative projects. (Linell et al, 2003: 559ff.) Elaborate multi-unit questions lead the interviewees' (IEs) to give longer, elaborated responses as in (1). Annotation symbols follow those by Gail Jefferson.

(1) [D03F0006: 0004-0042: 00003.970-00055.162] (simplified)
1IR: <ano> itsumo wa (.) kiite sore wo (0.3) bunseki shiteru hoo da to omou n desu ga,
"(I) suppose (you) usually listen to and analyze it [= the recordings], and"
2.hh kyoo jissai yatte doo deshita=
"how did (you) feel after actually doing [= participating in the recording] today?"
3IE: =.hh u:n nanka sonna koo:- kooyuu huu ni yatteru n dana: to omotta no to:
"Well (I) realized how (the recordings) are done, and"
4 ato wa: igai to muchuu ni natte iru jibun ni kizuki mashita ne((3 lines omitted))
"then (I) found myself unexpectedly enthusiastic"
5 kiite sore wo nante itteru ka tte yuuno wo bunseki suru hoo wo yatteru to:

"when (I) listen to and analyze what (participants) say"
6 wake wakannai koto wo yu- (yu/i)waretari suru to: hara ga tatsu n desu yone ((a line omitted))
"(I) get angry when (they) say nonsense thing"
7jibun de: iza hanashi wo suru to: shoo ga nai na tte
"when it's time to talk myself, (I think) it can't be helped"

In her multi-unit question, the IR contrasts IE's usual role as an analyzer (line 1) with her role on the day as a participant (line 2), and thereby imposes a complicated task on the IE: to tell how she felt about the prior recording while contrasting it with her usual feeling as an analyzer. In response to this question design, the IE contrasts how she felt at the scene (lines 3, 4 and 7) with what she usually thinks as an analyzer (lines 5-6), which comprises elaborated talk.

Trouble-anticipatory Questions

Another device for the IRs to elicit elaborated talk is to incorporate in questions the assumption that the IEs have some trouble. It often takes the form of multi-unit questions or question cascades.

(2) [D03M0037: 0006-0022: 00006.533-00030.159] (simplified)
1IR: doo deshi[ta (.)] tsu[kare mashi(ta)].hh]
"How was (the recording)? Did (you) get tired?"
2IE:[.hh][n::ma]a .hh]h soo desune: tsukareru tte yuu koto wa nai n desu
"Um well, it's not that (I) get tired"
3 kedo ne, ((a line omitted)) ma[a: so]nna ni (0.2) taihen dewa nakatta desu [kedo.]
"well, (the recording) was not very hard."
4IR: [hhh] [n : :] kinchoo
"Um, didn't
5 shimasen deshita ka?
(you) get nervous?"
6IE:<kin>choo wa <saisho> wa yappari:: shimasu ne. ano: .hhhh nante yuu n daroo nah
"At first (I) got nervous, as may be expected." "Well what should I say"
7>yappari shabe<ru koto sonna ni, (.) ano::hh tokui dato omotte nai n [de:]
"as may be expected, (I) don't think (I)'m so good at speaking"
8IR:[.hh]h <soo> desu ka:?
"Is it so?"

The IR asks the IE if he got tired in the recording. (line 1) When that assumption is denied (lines 2-3), the IR takes her turn to ask the second question: to ask if the IR got nervous in the recording. (lines 4-5) She manages to elicit an elaborated response (lines 6-7) by this second question, or rather by her persistent asking of trouble-anticipatory questions.

Exaggerated News Receipts

Exaggerated exclamations and assessments frequently appear in the IRs' responses to the IEs' talk. They enable the IRs to display that they regard the IEs' talk as newsworthy (Heritage 1984), which would encourage the IEs to talk more to elaborate.

Line 8 in (2) is an example of them. As is seen from annotation symbols, it is uttered in quite an exaggerated manner. The sound of *soo* is stretched, and almost the whole utterance is stressed.

It displays the IR's strong interest in the topic brought at line 7, which actually led to the extension of the topic.

Conclusion

This paper has described three linguistic devices utilized in a Japanese interview dialogue corpus to elicit elaborated talk: elaborate multi-unit questions, trouble-anticipatory questions and exaggerated news receipts. Through sequential analysis, the participants' orientation to longer and elaborated talk has been illustrated. This orientation reflects the corpus design for which they have to continue talking.

References

Heritage, J. C. 1984. A change-of-state token and aspects of its sequential placement. In Atkinson, J. M. and Heritage, J. C. (Eds), *Structures of Social Action: Studies in Conversation Analysis.* Cambridge: Cambridge University Press, 299-345.

Linell, P., Hofvendahl, J. and Lindholm, C. 2003. Multi-unit questions in institutional interactions:
Sequential organizations and communicative functions. *Text,* 23(4), 539-571.

Maekawa, K. 2002. Design, compilation, and preliminary analyses of the 'Corpus of Spontaneous Japanese.' *NTT-Stanford Workshop on Concept and Language Processing*, 13.

Using a Learner Corpus to Develop an EFL Grammar Teaching Curriculum

Susana Murcia Bielsa and Mick O'Donnell

Departamento de Filología Inglesa,
Universidad Autónoma de Madrid, Spain

Keywords**:** learner corpora, curriculum development, parsing, error analysis

Introduction

One of the aims of the TREACLE project is to determine which grammatical structures need to be taught to Spanish learners of English, at which proficiency level the structures are most critically taught, and how much attention should be given to each structure. To this end we have undertaken systematic manual and automatic analysis of a corpus of learner essays, each with associated proficiency level.

The corpus

This exploration of learner proficiency is based on manual and automatic analysis of 700,000 words of learner English produced by Spanish University students. Two separate corpora were included:

- The WriCLE corpus (Rollinson and Mendikoetxea, 2010), essays by students in an English degree, of which we use 560 essays, or 550,000 words;
- The UPV Learner Corpus (Andreu et al., 2010), shorter essays by students in English for Specific Purposes courses, of which we use 780 essays, or 150,000 words.

For both corpora, the writer's proficiency level was estimated by converting their score in the Quick Oxford Placement Test (UCLES, 2001) to a level in terms of the Common European Framework of Reference (CEFR, Council of Europe, 2001), being from lowest to highest: A1, A2, B1, B2, C1 and C2.

Profiling students through automatic syntactic analysis

To see what syntactic structures students at each proficiency level are capable of producing, the corpus was syntactically parsed, using the Stanford parser (Klein and Manning, 2003). From these parse trees, we extract out the set of syntactic features exhibited by each clause or phrase, e.g., relative-clause, passive-clause, modal-clause, wh-noun-phrase, etc. We can then measure the degree to which each of these syntactic features is used at each CEFR level, and by viewing the changing degree of use of each feature over the six levels, we can estimate where the structure would be most effectively taught to the student.

This talk will summarise our methodology for syntactic analysis, and results in regards to determining where particular grammatical structures should be taught.

Profiling students through manual error analysis

At the same time, we manually annotated a 75,000 word sub-corpus for errors, identifying 9,840 errors, each coded in terms of a scheme of 105 distinct error categories ranging over lexical errors, grammatical errors, pragmatic errors and register errors.

This study has, in our opinion, revealed various important results for curriculum design. Most importantly, 45% of the grammar errors produced by the students occur in the noun phrase, and almost half of these concern the wrongful inclusion or exclusion of the determiner. This strongly suggests that noun phrases in general, and determiners in particular, need to be given stronger emphasis in our grammar curricula.

While the syntactic analysis shows us the degree to which each syntactic structure is being used, the error analysis reveals the degree to which the student is struggling to master the production. This talk will explain how the error analysis extends upon the results gained from syntactic analysis, revealing how much emphasis needs to be assigned to each area of the grammar, based upon the frequency of errors the student make in that area.

Applicability of the results

While the work descibed here concerns Spanish learners of English, the outlined methodology could equally be applied to any other L1, given the proviso that the level of English is high enough for automatic parsing to be reliable. We found that the Stanford parser was accurate enough for the task where University learners were involved. Experiments with secondary learners in a Japanese context showed worse results.

Acknowledgements

The work here reported was funded by the Spanish Ministerio de Ciencia e Innovación, within the Project title: "Developing an annotated corpus of learner English for pedagogical applications" (FFI2009-14436/FILO), running from January 2010 to December 2012.

References

Andreu, M., Astor, A., Boquera, M., MacDonald, P., Montero B., and Pérez, C. 2010. Analysing EFL learner output in the MiLC project: An error it's*, but which tag?. In Campoy, M.C., Belles-Fortuno, B. and Gea-Valor, M.L. (Eds), *Corpus-Based Approaches to English Language Teaching*. London: Continuum, 167-179.

Klein, D. and Manning, C. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15* (NIPS 2002), Cambridge, MA: MIT Press, 3-10.

Council of Europe 2001. The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

UCLES 2001. *Quick Placement Test (Paper and pencil version)*. Oxford: Oxford University Press.

Rollinson, P. and Mendikoetxea, A. 2010. Learner corpora and second language acquisition: Introducing WriCLE. In Bueno Alonso, J. L., Gonzáliz Álvarez, D., Kirsten Torrado, U., Martínez Insua, A. E., Pérez-Guerra, J., Rama Martínez, E. And Rodríguez Vázquez, R. (Eds), *Analizar datos: Describir variación/Analysing data: Describing variation*. Vigo: Universidade de Vigo (Servizo de Publicacións), 1-12.

A Corpus-Based Analysis of College ESL Students' Argumentative Writing

Daehyeon Nam

Division of General Studies
Ulsan National Institute of Science and Technology, Republic of Korea

Keywords**:** ESL, argumentative writing, keyness

Introduction

The purpose of the current research is to examine argumentative writing written by a group of college students who use English as their second language in an academic setting. In the academic ESL setting, argumentative writing, in which the writers use arguments to support their opinion, is a type of academic writing taking a position for or against some statement on an issue (Hamp-Lyons, 1991). Drawing upon the recent attention to the importance of text in written academic discourse (Hinkel, 2003), the current research explores how Korean ESL college students provide their opinions on mass media, technology, and education. The current study investigates the grammatical words in the Korean students' argumentative writing to investigate the typical stylistic constructions (Bondi, 2010; Scott and Tribble, 2006). Argumentative writing is one of the frequently covered writing types in the commercial academic ESL writing textbooks (see Oshima and Hogue, 2006; Savage and Mayer, 2005). These textbooks might be useful to the ESL writers because the textbooks provide good layout of the argumentative writing by providing "templates" and "signal words". However, simply providing the templates and signal words may not be enough for the ESL writers to write argumentative writing effectively and to appropriately use certain words for expressing their opinions in the essay.

Keyness analysis is useful to extract information on what a text is written about and how the text is written. To investigate a written text, word frequency of the text may provide only a general overview or limited explanation of the text (Flowerdew, 2008; Laufer and Nation, 1999). Keyness analysis, however, provides useful information about the content of the text in a reliable way because it eliminates unnecessary function words and extracts content words by statistically comparing the text's word frequency list against a reference corpus's word frequency list (Scott and Tribble, 2006). In addition, the analysis can provide useful information on the stylistics of the text by investigating the grammatical keyword lists. Therefore, the current study employs the keyword analysis to capture the characteristics of the argumentative writing of Korean ESL students.

Methods and Analysis

The writing samples are collected in a voluntary manner, and the samples are analyzed and compared through "keyword analysis" (Scott and Tribble, 2006; Bondi and Scott (eds.), 2010). Argumentative writing samples on mass media, technology, and education were collected from

Korean undergraduate students at a large mid-western university. A total of 72 argumentative writing samples (24,558 running words) were collected. KeyWord program in the computerized concordancer software, WordSmith Tools 5 (Scott, 2008) was used to analyze the writing samples. KeyWords creates a list of keywords within a text by comparing its word frequency list against a large reference corpus. For the current analysis, the Michigan Corpus of Upper-Level Student Papers' (MICUSP) argumentative writing sample, a total of 563,705 words, was used as a reference. Since the keyword analysis was executed by comparing the frequency list of the text of interest and the analysis yields two types of keywords, either over-used words, i.e., positive keywords or under-used words, i.e., negative keywords.

Results

Of the 178 keywords of the argumentative writing samples (166 positive and 12 negative keywords), the following grammatical words were singled out from the list.

> The positive grammatical keywords are: *about; also; always; and; are; can; even; every; however; I; it; make; many; more; much; our; some; sometimes; their; there; they; too; we.* The negative grammatical keywords are: *an; as; between; he; him; his; in; of; this*.

The grammatical words may provide useful information on the stylistics of the Korean ESL writers' argumentative writing as Scott (2008) suggests.

Conclusion and Discussion

The keyword analysis of grammatical words provides insights of how Korean ESL students construct the argumentative writing styles. The keywords were referenced against the argumentative essays of MICUSP. This suggests that, although the two sets of the argumentative writings, ones from the Korean ESL students and others from students in the U.S., could be considered in the same writing type, it is also necessary to investigate how the two argumentative writings are stylistically different. Although the argumentative essays of MICUSP may not be considered to be a definite norm for academic ESL writing, the results of keyword analysis can serve as a diagnostic tool for evaluating ESL writing. In addition, the results can be resources for developing materials for the teaching academic writing to a certain group of students. Therefore, once the stylistic differences of the Korean ESL students' argument essays are identified, then the results can be used to teach ESL academic writing.

References

Bondi, M. 2010. Perspectives on keywords and keyness: An introduction. In Bondi, M. and Scott, M. (Eds.), *Keyness in texts. Am*sterdam: John Benjamins, 1- 18.

Bondi, M. and Scott, M. 2010. *Keyness in texts*. Amsterdam: John Benjamins.

Flowerdew, L. 2008. *Corpus-based analyses of the problem-solution pattern: A phraseological approach*. Amsterdam: John Behjamins.

Hamp-Lyons, L. 1991. Scoring procedures in ESL context. In Hamp-Lyons, L. (Ed.). *Assessing second language writing in academic context*. Norwood, NJ: Ablex Publishing.

Laufer, B. andNation, P. 1999. A vocabulary-size test of controlled productive ability. *Language Testing*, 16 (1), 33-51.

Michigan Corpus of Upper-level Student Papers. 2009. Ann Arbor, MI: The Regents of the University of Michigan.

Oshima, A. and Hogue, A. 2006. *Writing academic English* (4th ed.). White Plains, NY: Pearson Longman.

Savage, A. and Mayer, P. 2005. *Effective academic writing* 2: The short essay. Oxford: Oxford University Press.

Scott, M. 2008. WordSmith Tools 5.0 [Computer software]. Liverpool, UK: Lexical Analysis Software.

Scott, M. and Tribble, C. 2006. *Textual patterns: Keywords and corpus analysis in language education*. Amsterdam: John Benjamins.

How Much Input Do Learners Need? A Corpus-based Answer

Paul Nation

This study looks at how much input is needed to gain enough repetitions to learn the first 10,000 words of English. It uses corpora of various sizes and composition to see how many tokens of input would be needed to gain at least ten repetitions and to meet most of the words at each of the ten 1000 word family levels.
The corpus sizes needed range from just under 200,000 tokens at the 2nd 1000 word level to 3 million tokens at the 10th 1000 word level, or in terms of novels, from two to twenty-five novels (at 120,000 tokens per novel). Allowing for learning rates of between 1000 and 2000 word families a year, these are manageable amounts of input.

A New Wordtree Corpus Interface

Hilary Nesi

Department of English and Languages
Coventry University, UK

Keywords**:** interface; collocation, concordance, lexico-grammatical pattern, visualization

Introduction

The most used and cited corpora in the public domain tend to be those which enable the easiest and fastest extraction of relevant data. Potentially useful corpora are often underused because they lack online interfaces, or because their interfaces were originally designed for lexicographers and information scientists rather than for language teachers, language learners, and researchers in applied linguistics.

This paper reports on work in progress for a publicly funded project to develop a more usable and effectivevisual corpusinterface. It is currently being demonstrated as an interface to the BAWE corpus, utilising BAWE TEI-encoded header information to enable users to create and compare their own bespoke corpus subcomponents.

Existing text visualization tools

Interactive visualizations, such as those created by Clark (2008) and McCandless (2010), allow users to zoom in on specific details within large amounts of information, andencourage browsing because they are easy to navigate and are visually attractive. Doyle (2010) provides an overview of the visual displays in currentcorpus query tools. WordSmith Tools (Scott 2004), for example, can display collocation patterns and dispersion plots, and the BYU interface for COCA has interactive distribution charts. Such displays are limited by the fact that they arebased around tables and grids, however, and Doyle (2010: 161) argues thatmore effective visualisations can be achieved by transforming data into visual objects, for example in the form ofword cluster diagrams,also known as ‘word clouds’ (Clark, 2008), ‘tag clouds’ (Alex, undated) or ‘Wordles’ (Feinberg, undated). Word cluster diagrams group words within a text, using word size to indicate frequencyandpositioning and color schemes to indicate collocational relationships. They are generally used for decorative rather than for educational purposes, however. Te*xtArc* (Paley, 2009) adds interactivity to the word cluster diagram, using moving coloured lines to link words and collocations. Paley (2009) and Doyle (2010) claim that thisstimulates users to explore works of literature, butnevertheless theuse of *TextArc* as a corpus linguistic tool seemslimited because the lexical connections are not quantified, and there does not seem to be any facility to gather and store the data that is fleetingly revealed.

Word tree visualisations are essentially interactive Key Word in Context (KWIC) concordance lines grouped intotree-like branching structures. Although standard concordancing programs can make collocational patterns more apparent by sorting words preceding or following the search

term,it still takes skill to interpret concordance lines effectively, and novice concordance usersoften tend to scan the lines horizontally for meaning rather than vertically for lexico-grammaticalpatterns. Students quickly lose interest in classroom concordancing tasks, asThurston and Candlin (1998:278) point out.A tree structure provides the same amount of information on screen but reduces the amount of text; the tree is obviously not a piece of continuous prose, and the patterns are immediately apparent when the treeis read horizontally. Moreover the interactive'zoom in' feature makes word tree searches more visually exciting than concordancing.

Our project takes as its starting point theIBM *Many Eyes* Word Tree visualization (Wattenberg and Viégas, 2008). This works well, but is limited by the fact that text can only be uploaded in a single file of no more than 5MB, and the tree does not display any statistical information. Figure 1 shows a screenshot of a *Many Eyes* Word Tree for 'although', using the BAWE subcorpus of assignments belonging to the 'Critique' genre family (the entire BAWE corpus is too large for *Many Eyes*).

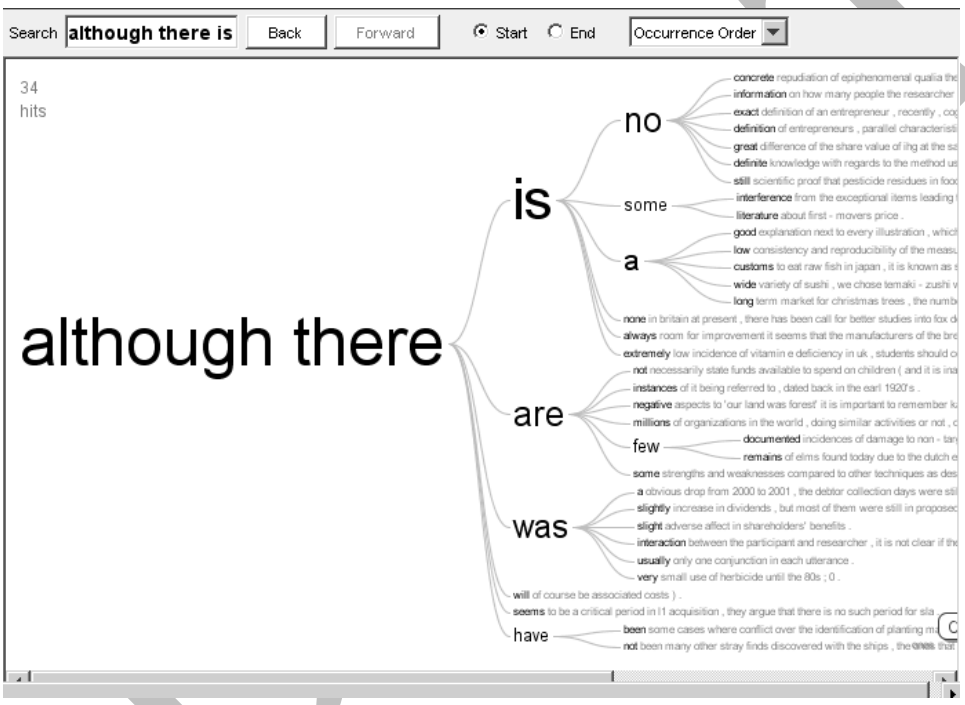


Figure 1: 'although' in a *Many Eyes* Word Tree

Our Word Tree specifications

Many Eyes Word Tree visualisations were shown to experienced corpus linguists, novice researchers, and language learners as a starting point for our project. We then set about developing a more flexible and elaborate interface in response to their feedback. Some of their main requirements for our new interface were increased data capacity, the facility to create subcorpora, word frequency, distribution and dispersion statistics (both raw and normalised), and access from the screen to the original text. This paper will demonstrate a beta version of ournew Word Tree, and discuss some of the issues surrounding its development. The blog for the project is http://cuba.coventry.ac.uk/wordtree.

References

Alex. undated. *Tagul*.http://tagul.com/

Clark, J. 2008. Neoformix: Discovering and illustrating patterns in data weblog. (see also http://www.neoformix.com)

Corpus of Contemporary American English (COCA). http://corpus.byu.edu/coca

Doyle, P. 2010. Viewing language patterns: data visualisation for data-driven language learning. In Ho, C. M. L., Anderson, K. and Leong, A. P. (Eds), Transforming Literacies and Language: Multimodality and Literacy in the New Media Age. London, Continuum, 149-166.

Feinberg, J. undated. *Wordle*.http://www.wordle.net/

Many Eyes. Visual Communication Lab, Collaborative User Experience Research Group, IBM. http://www-958.ibm.com

McCandless, D. 2010. *Information is Beautiful.* London: HarperCollins. (see also http://www.informationisbeautiful.net/)

Paley, W. 2009. Interface and mind. *Information Technology,* 15 (3), 131-141

Paley, W. undated. *TextArc*.http://textarc.org/

Scott, M. 2004. *WordSmith Tools* version 4. Oxford: Oxford University Press.

Thurstun, J. and Candlin, C. 1998. Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes,* 17, 267–280.

Wattenberg, M. and Viégas, F. B. 2008. The Word Tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions*, 14 (6), 1221-1228.

An Account of the English Loanwords in Singapore Chinese – A Corpus-Based Approach

Serene S.H. Ng

Singapore Centre for Chinese Language
Nanyang Technological University, Singapore

Keywords**:** loanword, Singapore, Chinese, borrowing, corpus

1. Introduction

Loanwords refer to words adopted by speakers of one language from a foreign language.A loanword is also known as a borrowing. Borrowing usually occurs when there are lexical gaps for concepts, thoughts, and inventions as well as issues of prestige and modernity. Due to the British rule of Singapore during the colonial time, English has become the lingua franca of Singapore and was made the main language of administration upon Singapore independency. Since 1981, Singapore Ministry of Education implements English as the medium of instruction in schools and six years later, it fully rolled out its bilingual education system (Cheah, 2003). Today, English remains the main language of policy, administration, economic and education in Singapore. With a population of a majority Chinese (76%), Singapore's English-dominant environment leads to the absorption of English words into the Chinese lexical system. On this basis, this study seeks to investigate the English loanwords found in Singapore Chinese.

2. Methodology

This study was carried out with the construction of a specialised corpus which collected data from the recent 50 years Singapore Chinese written texts. In order to ensure that the study is representative, data were drawn from local Chinese newspapers and locally published Chinese books. In total, the corpus collected about 5.2-million Chinese character-tokens from the Singapore Chinese written text from 1960 to 2009. Random sampling was adopted during the data collection process so as to ensure a balanced and unbiased corpus. For the books data, they were collected from literature, bibliographies, historical records, education, economics, law as well as medical, which accounts for 64% of the main corpus. For the newspapers, data were collected from headline news, social news, regional and world news, entertainment and leisure information as well as articles, which accounts for 36% of the main corpus.

3. Results

Findings reveal that English loanwords found in Singapore Chinese mainly take the form of transliteration, translation, as well as a combination of both. Among the various forms, transliteration is a more common means. Investigation found that the English loanwords in Singapore Chinese can be classified into few broad categories, namely transportation, social

phenomena, trends and others.

3.1 Transportation

The inventions of most of the means of transport were from the West resulted in the borrowing of the related terms into Singapore Chinese.‘Bus’ was transliterated as ‘bāshì’ (巴士) while ‘lorry’ is known as ‘luólí’ (罗厘).‘Taxi’ was transliterated as ‘déshì’ (德士), whereas ‘pick-up truck’ is known as ‘bìjiǎ’ (必甲).

‘Motorcycle’, ‘Jumbo jet’ and ‘Airbus’, however, were absorbed through a combination of transliteration and translation.‘Motor’ was transliterated as móduō (摩哆), while ‘cycle’ was translated as ‘chē’ (车). Thus ‘motorcycle’ is known as ‘móduōchē’ (摩哆车).‘Jumbo’ was transliterated as ‘zhēnbǎo’ (珍宝), while ‘jet’ was translated as ‘jī’ (机). Hence, ‘jumbo jet’ is known as ‘zhēnbǎojī’ (珍宝机) in Singapore.‘Air’ was translated as ‘kōngzhōng’ (空中), while ‘bus’ was transliterated as ‘bāshì’ (巴士). Thus ‘Airbus’ is known as ‘kōngzhōng bāshì’ (空中巴士) in Singapore Chinese.

3.2 Social Phenomena

It is interesting to note that some English loanwords found in Singapore Chinese are related to social phenomena. For instance, generation gap, junk mails, the litterbugs who create environmental problems, the road bully who create problems on the road and the street gangs who cause social problems. Findings reveal that all these social phenomena words were borrowed through direct translation.‘Generation gap’ was translated as ‘dàigōu’ (代沟), while ‘road bully’ is known as ‘lùbà’ (路霸). ‘Street gang’ was referred to as ‘jiētóudǎng’ (街头党).‘Junk mail’ is known as ‘lājī yóujiàn’ (垃圾邮件), whereas ‘litterbug’ was translated as ‘lājīchóng’ (垃圾虫). In Chinese, ‘lājīchóng’ (垃圾) refers to garbage, trash or waste.

3.3 Trends

Findings show that some trends which were originated from the West were being borrowed into Singapore Chinese lexical system.‘Mini-skirt’ and ‘bungee jump’ were being borrowed through a combination of transliteration and translation.‘Mini’ was transliterated as ‘mínǐ’ (迷你), while ‘skirt’ was translated as ‘qún’ (裙). Hence ‘mini-skirt’ is known as ‘mínǐqún’ (迷你裙).‘Bungee’ was transliterated as ‘bǎngjǐn’ (绑紧) while ‘jump was translated as ‘tiào’ (跳). Thus ‘bungee jump’ is known as ‘bǎngjǐntiào’ (绑紧跳) in Singapore.‘Catwalk’, ‘charity walk’ and ‘hot-pants’ on the other hand, were borrowed through direct translation.‘Catwalk’ was translated as ‘māobù’ (猫步) while ‘charity walk’ was translated as ‘yìzǒu’ (义走).‘Hot-pants’ was translated as ‘rèkù’ (热裤).

3.4 Others

Other transliterated English loanwords include ‘percent’, known as ‘bāxiān’ (巴仙); ‘boycott’, known as ‘bēigě’ (杯葛); ‘coupon’, known as ‘gùběn’ (固本); ‘bookie’, known as ‘bǔjī’ (卜基); ‘tips’, known as ‘tiēshì’ (贴士); ‘cushion’ known as ‘gǔshēn’ (古申); ‘license’ known as ‘lǐshēn’ (礼申) and ‘mosaic’ known as ‘máoshí’ (毛石). Some other English loanwords found in Singapore Chinese were absorbed through translation. These include ‘drinking water’, known as ‘shíshuǐ’ (食水); ‘cash card’, known as ‘xiànjīnkǎ’ (现金卡);

'road hump', known as 'lùlóng' (路隆)　and 'yellow page', known as 'huángyè' (黄页).

4. Conclusion

In conclusion, the English loanwords found in Singapore Chinese reflect the cultural exchange between the east and the west. From a micro viewpoint, words borrowing occurs in association with changes in Singapore society at different period of time, while from a macro perspective, loanwords found in Singapore Chinese take place in association with changes in the globe.

References

Cheah,C.M. 2003. *Jiaoxue yu ceshi* [Teaching and assessment]. Singapore: Singapore Chinese Teachers' Union.

Chen, C.Y. 1993. *Aspects of Mandarin Chinese*. Singapore: Chinese Language & Research Centre, National University of Singapore.

Chew, C.H. 2002. An outline of variations of Singapore Mandarin. *Zhongguo Yuwen*, *6,* 508-514.

Chew, C.H. 2008. The variations of Singapore Mandarin and the strategies in dealing with the variations. In *Teaching and learning of Chinese language Series. Vol 5*. Singapore: EPB Pan Pacific.

Langacker, R.W. 1968. *Language and its structure*. New York: Harecourt, Brace & World, Inc.

Ng, S. 2010. Using corpus to investigate the lexicon variants of Singapore Mandarin. *Corpus, ICT, and Language Education*. UK: University of Strathclyde Publishing

Ngom, F. 2002. Linguistic borrowing as evidence of the social history of the Senegalese speech community. *International Journal of Sociology of Language*, 158, 37-51.

Interlanguage Variation in Past Tense Marking in Japanese EFL Learners' Spoken and Written Corpora

Mariko Nomura

Graduate School of Area and Culture Studies
Tokyo University of Foreign Studies, Tokyo, Japan

Keywords**:** aspect hypothesis, discourse hypothesis, narratives, spoken and written corpora

Introduction

This study investigates the distribution of interlanguage past tense verb forms in Japanese EFL learners' spoken and written language from two perspectives: lexical aspect and discourse structure. Previous SLA studies have suggested that second language learners' use of past morphology is influenced by the inherent semantic aspect of verbs (i.e. aspect hypothesis) (Bardovi-Harlig and Reynolds, 1995; Bardovi-Harlig, 1998; Kumpf, 1984) and the narrative structure (i.e. discourse hypothesis) (Bardovi-Harlig, 1995, 1998; Kumpf, 1984). There are only a few studies which have tested these hypotheses, using both oral and written data from the same learners. No research has been conducted on how both hypotheses interact with respect to Japanese EFL learners' past reference using oral and written discourse pairs. The present study attempts to examine how the distribution of simple past is influenced by lexical aspect and narrative structure using narratives produced by Japanese secondary school students.

Background

Bardovi-Harlig (1998) suggests that both the aspect hypothesis and the discourse hypothesis are necessary to account for the distribution of past morphology in interlanguage. The results of her study showed that lexical aspect and narrative structure conspired to shape the distribution of tense-aspect morphology in interlanguage (p. 501).

Lexical, or inherent, aspect is most known as four categories proposed by Vendler (1967); states, activities, accomplishments, and achievements. The aspect hypothesis predicts that telic verbs (i.e. accomplishments and achievements) are more likely marked by simple past morphology (Bardovi-Harlig, 1998). Narrative discourse is comprised of two parts; the foreground and the background. The foreground is the language of the actual story line, that is, "the parts of the narrative which relates events belonging to the skeletal structure of the discourse", while the background is "the language of supportive material which does not itself narrate the main events" (Hopper, 1979:213). The discourse hypothesis predicts that learners use emerging verbal morphology to distinguish foreground from background in narratives; it predicts that verbs in the foreground tend to be marked by simple past morphology (Bardovi-Harlig, 1998).

Research Questions

The following two research questions are addressed in the present study.

RQ1: Do the inherent meaning of verbs and narrative structure influence the distribution of Japanese EFL learners' past tense use?

RQ2: Do the aspect hypothesis and the discourse hypothesis interact with respect to learners' past tense reference?

Method

The data used in this study was a subcorpus of spoken and written corpora comprising 324 Japanese secondary school students' discourse pairs on three topics. Spoken and written discourse pairs on the topic "Impressive school events" were used as narratives for the investigation of learners' past tense use. The tokens of spoken and written data totaled 4,023 and 5,360 English words respectively.

In order to answer RQ1, each verb used in past-time contexts was coded into one of the four aspectual classes proposed by Vendler (1967); states (STA), activities (ACT), accomplishments (ACC), and achievements (ACH). Subsequently, the distribution of the past tense use of these four classes was examined in each production mode. In determining inherent lexical aspect, the tests used by Shirai and Anderson (1995) were adopted.

As for narrative structure, sentences or clauses in the discourse were classified into foreground (F) and background (B). After that, the simple past use was examined in each grounding in spoken and written discourse. In addition, the distribution of base/present forms was examined. The distribution of lexical and *be* verbs was analyzed separately.

In order to answer RQ2, the interaction of the aspect hypothesis and the discourse hypothesis was examined by combining the aspect and discourse analyses in RQ1.

Results

1. The influence of lexical aspect on past tense use

The distribution of simple past across aspectual class in both the spoken and written narratives supported the prediction of the aspect hypothesis; achievements show the highest rate of simple past inflection. Overall, the rates of appropriate simple past use in the written narratives were higher than those in the spoken narratives, suggesting that verbal morphology is more likely to be used in written mode with more monitoring effect.

2. The influence of narrative structure on past tense use

The distribution of simple past by grounding in both the spoken and written narratives supported the prediction of the discourse hypothesis; both modes of narratives showed greater use of simple past with lexical verbs in the foreground than in the background. *Be* verbs in the past tense were used more dominantly in the background than in the foreground in both modes.

3. The interaction of the aspect hypothesis and the discourse hypothesis

The distribution of simple past in each aspectual class by grounding in the spoken and written narratives suggests that lexical aspect and discourse structure have an interacting effect on the

variable use of past tense marking in learner interlanguage.

References

Bardovi-Harlig, K. 1995. A narrative perspective on the development of the tense/aspect system in second language acquisition. *Studies in Second Language Acquisition*, 17, 263-291.

Bardovi-Harlig, K. 1998. Narrative structure and lexical aspect: Conspiring factors in second language acquisition of tense-aspect morphology. *Studies in Second Language Acquisition,* 20, 471-508.

Bardovi-Harlig, K. and Reynolds, D. W. 1995. The role of lexical aspect in the acquisition of tense and aspect. *TESOL Quarterly*, 29, 107-131.

Hopper, P.J. 1979. Aspect and foregrounding in discourse. In Givón, T. (Ed.), *Syntax and Semantics: Discourse and Syntax*. New York: Academic Press, 213-241.

Kumpf, L. 1984. Temporal systems and universality in interlanguage: A case study . In Eckman, F., Bell, L.and Nelson, D. (Eds), *Universals of Second Language Acquisition*. Rowley, MA: Newbury House, 132-143.

Shirai, Y. and Anderson, R.W. 1995. The acquisition of tense-aspect morphology. *Language*, 71, 743-762.

Vendler, Z. 1967. Verbs and times. In Vendler, Z. (Ed.), *Linguistics and Philosophy*. Ithaca, NY: Cornell University Press, 97-121.

Historical Analyses of English Negative Affixes from Middle English to Present-day English

Akira Okada

Department of English, Daito Bunka University,
Tokyo, Japan

Keywords**:** negative prefixes, frequency, obsolete words, BNC, *OED*

Introduction

This paper aims at describing how English negative affixes have been used from the Middle English period (ME) to Present-day English (PE) by native speakers of English. It is true that English has some negative affixes in the modern usage, all of which are said to have survived as relatively productive morphemes in spite of the fact that they share similar meanings. Given that, linguistically speaking, similar morphemes exist in English, it is discernable that two or more distinctive morphemes occur with regard to words as word pairs, such as is the case of *in-* / *un-* *certain* or *in-* / *un-* /*dis-honest*. Moreover, it is interesting to investigate how negative morphemes such as *un-*, *in-*, *-less* and so on, have been used during the course of the long history of English. In order to recognize how many such pairs can be found in the modern usage, the British National Corpus (BNC) will be of great use. *The Oxford English Dictionary* online (*OED*) will also supply necessary information for investigating how negative morphemes, especially, doublets have appeared in the past texts. There are two points in this paper to examine the features of negative prefixed words as follows:

(1) Frequency in PE

(2) Obsolete words or rare usages in PE

Method

We will deal with the English negative affixes which exist in PE like *un-, in-, non-* and *-less*. At first, with the research of BNC the top twenty words attached by each of the affixes above will be presented. Secondly, let us analyze the frequency of occurrence. Thirdly, the past examples shown in the *OED* will be illustrated. Lastly, we will consider the process of how negative prefixes have been used from ME to PE.

Definitions

There are some definitions in order to make this paper clear.

(1) Regard *il-, im-* and *ir-* as allomorphs of *in-*.

(2) Reversal *un-* which is attached to verbs is not dealt with.

(3) For the supplement of the *OED*, *the Middle English Dictionary* (*MED*) is often used.

(4) In addition to the affixes concerned here, some other negative prefixes like *dis-, a-* and so on will also appear for the comparison.

References

Allen, M. 1978a. *Morphological Investigations*. Ph.D. Dissertation. Connecticut: The University of Connecticut.

---- 1978b. *The Morphology of Negative Prefixes in English*. Papers from Annual Meeting, North Eastern Linguistics Society 8th. 1-11.

Bauer, L. 2001. *Morphological Productivity*. Cambridge: Cambridge University.

Jespersen, O. 1917. *Negation in English and Other Languages*. KØbenhavn: Bianco Lunos Bogtrykkeri.

----- 1954. *A Modern English Grammar on Historical Principles: Part5 Sounds and Spellings*: George Allen and Unwin Ltd.

----- 1954.*A Modern English Grammar on Historical Principles: Part VI, Morphology*. London: George Allen and Unwin.

Kwon, H. S. 1997. "Negative prefixation from 1300 to1800: A case study in in-/un- variation." *ICAME Journal, Computers in English Linguistics No.21* [On-line] (24 June 1997). Available from URL < http://icame.uib.no/ij21/>

Marchand, H. 1969. The *Categories and Types of Present-day English Word-formation*. Wiesbaden: Otto Harrassowitz.

Plag, I. 2009. *Word-Formation in English*. Cambridge: Cambridge University Press.

Lieber, R. 1990. On *the Organization of the Lexicon*. New York: Garland.

-----2005. Mo*rphology and Lexical Semantics*. Cambridge: Cambridge University Press.

-----2010. In*troducing Morphology*. Cambridge: Cambridge University Press.

Anaphoric Demonstratives: a Cross-linguistic Corpus-based Approach

Marco Rocha

Departmento de Língua e Literatura Vernáculas
Universidade Federal de Santa Catarina, Brazil

Keywords**:** corpus-based translation studies, cross-linguistic approaches, anaphora, demonstratives

Introduction

The paper presents a corpus-based study on anaphoric demonstratives (ADs) in translations of English into Chinese. The study is part of a wider investigation on ADs which also includes Brazilian Portuguese.

The aim is to establish monolingual patterns of usage as well as patterns of translation. As pointed out in Johansson (2007), cross-linguistic studies may lead to insights regarding one of the languages included in the study that could not be obtained through its study in isolation. According to Diessel (2003), demonstratives are a linguistic item with unique features that belong to the basic vocabulary of languages. Although classified as pronouns in traditional grammars, they are often used as determiners as well in many languages. However, it may be said that demonstratives are strongly associated to pointing in their exophoric usage. This may explain why they are so old as to make it impossible to carry out etymological analysis of these items in a number of languages. The origins of these words would be associated to a very basic act of drawing another person's attention to elements of the situation involved in the ongoing interaction. The analysis of tokens of demonstratives in authentic discourse reveals their versatile nature, identifying uses as anaphors that refer to noun phrases, to other pronouns and to chunks of preceding text as well as to implicit antecedents suggested by clauses or discourse chunks. In this study, an AD is understood as a demonstrative that is not a determiner. Syntactically, the AD is the head of a noun phrase.

Methodology

The Babel Corpus was used as a source of data. A proportional sample of 199 tokens of ADs was analysed, including tokens of *this* (104), *that* (71), *these* (17) and *those* (7) used as anaphors. Categories to classify the translations were created. The correspondent ADs in Chinese *这* and *那*, as well as the plural forms *这些* and *那些*, were respectively treated as the expected translations and made up one category.

Translations as two-character combinations of *这* and *那* which were tagged as pronouns in the Babel Corpus make up the second category in the classification scheme. The third category groups tokens of ADs translated as anaphoric noun phrases. Renderings in which the retextualisation removes the AD were classified as omissions, since the text is reorganised to such an extent that it is not possible to identifiy a specific token as a translation of the English AD. Translation pairs in

which the AD was replaced by a Chinese pronoun other than the expected forms or the two-character combinations were classified as substitutions. The study attempts to uncover text patterns associated to these choices in target language texts by analysing contextual features of the source language text.

Results

A considerable variety of unexpected renderings was observed. Within the expected-translation category, *that* was rendered as *这* in 24 cases, breaking with the proximal-distal correspondence, and as *这些* in one case, which breaks also with the singular-plural correspondence. The AD *these* was translated as the singular form *这* in 3 cases. There were 18 tokens (9.04%) of translations as two-character combinations of *这* and *那*. The dominant combination is *这样* (14 tokens), and there is also one token of the distal *那样*. There were 41 translations of English DAs as nonpronominal anaphoric noun phrases (20.60%). There were 28 cases of omission (14.07%). Substitutions amounted to 15 cases (7.54%). Table 1 below summarises the results.

Table 1 - English ADs by Chinese translations in the Babel Corpus

Chinese tr.	expected 这-这些-那-那些				2-ch. comb.				NP	omiss.	subst.	Total
English ADs	□	□些	那	那些	□么	□□	□里	那么				
this	54	0	0	0	1	13	1	0	22	7	6	104
that	24	1	10	0	1	1	0	1	10	17	6	71
these	3	1	0	0	0	0	0	0	6	4	3	17
those	0	0	0	3	0	0	0	0	3	1	0	7
Total	82	2	10	3	2	14	1	1	41	28	15	199

The most frequent AD in English, *this*, is translated as the most frequent AD in Chinese *这* in 54 cases. The correspondence between singular proximal ADs holds reliably for *this* within the expected translation category. Patterns for the most common 2-character combination 这样 include English anaphoric DAs used as prepositional objects in phrases such as *like this* and *for this*; subjects of future-tense verb forms *this will mean* and t*his will make possible*, typically translated as the fixed expression *这样一来*; tokens of the conditional clause *if this were really so*; tokens of what-clause constructions such as *this is what*...; and objects of verbs *do*, *think* and *believe*. The token translated as *这里* has a locality as referent. Tokens of *this* and *that* translated as *这么* and *那么* seem to be variations of the 这样 patterns and a translator's choice. The 24 cases of *that* translated as *这* (35.21% of all *that* tokens) contain 21 tokens which are subjects of present tense sentences in the English original, often as part of fixed phrases such as *that's why*..., translated regularly as *这 就 是*, *that means*... (*这 就 意味着*), or, simply, *that is*... (rendered as *这* and predicative with the copular verb omitted). Contrastively, seven of the ten *that* tokens translated as *那* were subjects of sentences in the past tense in the English original. Translators seem to feel thus that the notion of tense in English may be expressed in Chinese by means of a

more strict proximal-distal contrast in the use of ADs, often influenced by the notion of perfective aspect or completed action.

A substantial share of the English ADs translated as anaphoric noun phrases are either generic or superordinate noun phrases, such as *这种 看法* or *这 件 事*. These specific decisions seem in fact to signal that the use of ADs to refer to passages of text is more restrictive in Chinese. Thus, preferred translations for sentences such as *Can that happen in China?* seem to be *中国 会 有 这 事情 吗*？, avoiding the reference by means of an AD.

Omissions also seem to occur because the expected forms of translation require the removal of the AD, such as the translation of *that's right* as *是的* or the rendering of *that 's OK* as *没关系*.. Pronouns commonly used as translations of English ADs in cases classified as substitutions include *之*, *此* and *其*. The influence of fixed phrases and idioms is also strong over this category of renderings, pointing to solutions such as *因此* to translate *because of that* and *此时此刻* to translate *this was the moment*. It seems possible eventually to map such collocations with a substantial degree of reliability. The combination of tense, aspect, expected forms of textual reference and collocations may thus constitute useful knowledge for the improvement of machine translation between English and Chinese.

Acknowledgment

This research project is funded by the Coordenadoria de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES) through the grant BEX 6566-10-3, for which I am deeply grateful.

References

Johansson, S. 2007. Seeing through multilingual corpus. In Facchinetti, R. (Ed.), *Corpus linguistics25 years on*. Amsterdam: Rodopi.

Diessel, H. 2003. *Demonstratives in language use and grammar*. Handout, San Marino Summer School. Max Planck Institute for Evolutionary Anthropology.

Construction of a Large-scale Translation Corpus and its Research and Pedagogical Implications

Yasumasa Someya[1], Atsuko Kikuchi[1], Shiro Akasegawa[2], and Yoichi Yamaoka[3]

[1]*Department of Foreign Language Studies, Kansai University, Osaka, Japan*

[2] *Lago Institute, Shiga, Japan.*

[3] *Freelance Translator,Tokyo, Japan.**

Keywords**:** translation corpus, LWP (Lago Word Profiler), Japanese compound verbs

We are currently constructing a large-scale Japanese-English/English-Japanese translation corpus consisting of both fictional and non-fictional works of various kinds. As of this writing, we have approximately 20 million words in the English part of the corpus alone. Text alignment work will be commenced by the end of this year, which is followed by a various types of annotation work -- all of which require a labor intensive manual work. Although the corpus itself is still far from complete, we now have a working beta version of our online corpus analysis tool, which we call LWP (Lago Word Profiler, see Fig 1.).

In this paper, we will first discuss the purpose and scope, as well as the limitations of the current project. This will be followed by a brief discussion as to the status of translation as linguistic data in the main stream linguistics. Our position is that translation corpora are in fact a valuable and very much reliable resource for linguistic inquiries not only in translation studies, but also in the areas of contrastive analysis, intercultural communication and second language acquisition. The main feature of translation corpora (aka, parallel corpora) is that they consist of source language texts on the one hand and target language texts on the other. In a monolingual corpus, words and/or expressions contained therein are treated something as given, or as unanalyzed entities; their "meanings" are left for each researcher to interpret. The corpus, in other words, has basically nothing to say about the meanings of the language in question. Translation corpora, on the other hand, deal directly with the meanings, both overt and covert, of a particular lexical item or combinations thereof in the source language and how these meanings are transferred into the target language. We therefore believe that by closely observing and describing what happens when something is translated, we may be able to have a better and informed understanding of the so-called linguistic intuition that a native speaker of a particular language is assumed to possess. If the translator is a native speaker of the target language (as in when a native speaker of English translates a Japanese novel into English), his or her translation will be reflecting not only the linguistic preference of the users of that language but also the underlying socio-cultural norms commonly shared by them. Thus, we agree with Mauranen (2002) when she says that "Parallel corpora can serve as a useful starting point for cross-linguistic contrasts because findings based on parallel corpora invite further research with monolingual corpora in both languages."We also believe that translation corpora offer one of the best available resources for intercultural

communication studies and foreign language pedagogy.

In the latter half this presentation, we hope to show that we have good reasons to believe so by providing some examples taken from a Web-based experimental module of the LWP. The research question we will be using for this demonstration will be: What is the exact meaning of the Japanese compound verb [$_{V1}$ XX [$_{V2}$ KIRU]] (as in "Ii-KIRU" or "YARI-KIRU"), and how does it differ from a similar compound verb [$_{V1}$ XX [$_{V2}$ TSUKUSU]] (as in "Ii-TSUKUSU" or "YARI-TSUKUSU")?

Fig 1 shows a sample image of the LWP User Interface. The pane of the left shows all the possible lexical patterns of the search string (Fig 1 shows search results of the verb TATERU, or "to set up"). The middle pane lists NP+VP collocations with statistical data. On the right pane are shown all the actual sentences that contain the target string in both original and translated texts.

Fig. 1 Sample image of the LWP User Interface

The results of our analysis generally concur with those of previous research (eg. Sugimura 2008, Niinuma 2010, Lee 1997) as to the possible meaning(s) of these two compounds, but revealed some very interesting conceptual components in the XX-KIRU compound that has not been discussed in the previous literature. We also found that in the V1 slot of the [$_{V1}$ XX [$_{V2}$ KIRU]] compound comes an action-oriented verb, whereas the V1 slot of the [$_{V1}$ XX [$_{V2}$ TSUKUSU]] compound is generally occupied by an object-oriented verb because of their respective meaning properties, although there are exceptions. Further details of the analysis as well as the research and pedagogical implications of the proposed translation corpus and the LWP will be discussed in our presentaion.

References

Hunston, S. 2002. *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Laviosa-Braith-Waite, S. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*, PhD Thesis, Manchester, UMIST.

McEnery, A. M. and Xiao, R. Z. 2007. *Parallel and comparable corpora: What are they up to?*Incorporating Corpora: Translation and the Linguist. Translating Europe . Multilingual Matters, Clevedon, UK.

Mauranen, A.2002. Will 'translationese' ruin a contrastivestudy? *Languages in Contrast,* 2(2), 161-86.

Olohan, M. 2004. *Introducing Corpora in Translation Studies.* Abingdon: Routledge.

李暻洙1997. 「中間的複合動詞「きる」の意味用法の記述――本動詞「切る」と前項動詞「切る」:後項動詞「切る」と関連づけて」『世界の日本語教育７』国際交流基金日本語国際センター

新沼史和2010. 「複合動詞「ー切る」に関する統語論的分析」『高知学園短期大学紀要』第40号：23-32.

杉村泰2008a. 「複合動詞「ー切る」の意味について」『言語文化研究叢書』7名古屋大学大学院国際言語文化研究科.

* Mr. Yamaoka, one of our project members and a renowned translator with more than 70 major translation works in the past 30 years, suddenly passed away on August 20, 2011. Taking this opportunity we would like to acknowledge his valuable contribution to the current project and express our deepest sympathy to the bereaved family. May his soul rest in peace.

Grammar of *HAN* 'say; said' in Sakizaya: a Corpus-based Perspective

Li-May Sung

Graduate Institute of Linguistics
National Taiwan University, Taiwan

Keywords**:** Sakizaya, Austronesian (Formosan) language, verbs of saying, grammaticalization

Sakizaya is an Austronesian (Formosan) language spoken in the east coast of Taiwan. The present paper aims to investigate how the speakers of Sakizaya encode evidentiality into their language. We will focus on *han* in this research, based mainly on naturalistic data taken from the *NTU Corpus of Formosan Languages* totaling 23 narratives with approximately 114 minutes, consisting of 1 daily conversation and 22 narratives of *Pear* stories, *Frog* stories, folktales, legends and daily life, which run approximately 113 minutes with a total of 2673 IUs. Both core sense and (extended) functions of *han* are carefully examined in order to comprehend how Sakizaya speakers display their own stance toward the information they are presenting.

In addition to its general, well-described function as a marker of focus/voice in (1) (cf. Shen 2008, Wu 2006), many instances of *han* mainly mark a certain kind of evidentiality shown in (2), extended from the semantic feature of the speech verb 'say/call' to signal stance-taking, enabling a speaker to comment on the source of the current utterance. In other words, they are packaged particularly with certain subjective speaker attitudes toward the action event indicating source/(in)directness of knowledge. Similar grammatical devices are found in Plains Cree and Philippine languages and termed as impersonal passives or unspecified subject constructions in Siewierska (1984), Shibatani (1988), and Dryer (1997). The subject has a meaning roughly paraphrasable as 'by someone/something' or 'by they/people (generic)'.

In addition, *han* also appears in the scenario where the propositional content is perceived through a direct visual perception of the narrator. This is exemplified in (3). The core sense of 'say' is fading away in this usage. Furthermore, the speech verb of saying *han* seems closely associated with a reinterpretation of "reason" and "conditional mood" as shown in (4)-(5), one of the stages in a grammaticalization hierarchy attested in many languages that verbs of saying often follow. (cf. Romaine and Lange, 1991; Saxena, 1995; Güldemann and von Roncador, 2002)

In (6), *han* may be further bleached and grammaticalized into a pragmatic discourse marker of speakers "calling" the attention (of hearers) on different characters in the process of story-telling. We will then conclude our study with a possible grammaticalization pathway of the foregoing functions of *han* in Sakizaya.

(1) SkzyNr-TYT_frog

4... sa kina wawa sa,\

DM this.NOM child DM

5... patepazeng **<u>han</u>**niza kyabalaut nuni-pa-habay-an niza,\

PATE put HAN 3SG.GEN this.NOM frogGEN NI-CAU-millet-NMZ 3SG.GEN

6... i tini i kuleng.\

LOC there LOC jug

'The child put the frog in the millet jug.' (Lit: The frog was put by the child in the millet jug.)

(2) SkzyNr_YAM_sapad

40... (0.7) u sapadien **<u>han</u>** ku nu luwak ani-pa-ngangan saka.\

CN place.name HAN Nom Gen Chinese ANI-Caus-name SAK

'Chinese named/called it as Sapadien.'

(3) SkzyNr_YAM_frog

86...(3.6) ma-htik ciniza i tapad

AF-fall.off 3SG.NOM LOC cliff

melaw <u>han</u> sa u nanum kya ka-htik-an niza.\

see HAN DM CN water that.NOM KA-fall.off-LF 3SG.GEN

'He fell off the cliff -- (I, the narrator) saw it happen -- and he fell into water.'

(4) SkzyNr-Palamal_Kaniw

49. **<u>zay han</u>**

because

50. uyni ulamal haw ku

thisCN fireINT NOM

51. saka tebek a ma-patay

therefore fall LNKAF-dead

'Because of this fire, (Sakizaya) was destroyed.'

(5) SkzyNr_LWY_festival

61.. ma-lecan mi-ki-tutung-ay o== **anu <u>han</u>** nu._

AF-same AF-KI-call-AY INT COND say.so.PF GEN

62.. tamdaw u sa-patubeli ku buyoh./

person CN SA- respond NOM mountain

'It's like that a person shouts "O" and (as if) another person seems to respond in the mountain.'

(6) SkzyNr-TYT_frog

24... ya wawa **<u>han</u>** tu,\

YA child HAN PFV

25... buhat satu tunian u,\

open SATU this CN

26... sa sasaedeb sa,\

DM door DM

'The child opened the door.' (Lit: As for the child, (he) opened the door.)

References

Dryer, M. 1997. Passive vs. indefinite actor constructions in Plains Cree. In Pentland, D. (Ed), *Papers of the Twenty-Sixth Algonquian Conference*. Winnipeg: University of Manitoba.

Güldemann, T. and von Roncador, M. (eds.). 2002. *Reported Discourse: A Meeting Ground for Different Linguistic Domains*. Typological Studies in Language 52. Amsterdam: John Benjamins.

Romaine, S and Lange, D. 1991. The use of *like* as a marker of reported speech and thought: a case of grammaticalization in progress. *American Speech,* 66, 227-279.

Saxena, A. 1995. Unidirectional grammaticalization: diachronic and cross-linguistic evidence. *Sprachtypologie und Universalienforschung* 48, 350-372.

Shen, W.C. 2008. *Sakizaya Syntax: With Special Reference to Negative, Interrogative and Causative Constructions*. MA thesis. National Taiwan University.

Shibatani, M. 1988. Voice in Philippine languages. In Shibatani, M. (Ed), *Passive and Voice*. Typological Studies in Language 16. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Siewierska, A. 1984. *The Passive: A Comparative Linguistic Analysis*. London: Croom Helm.

Wu, J.J.L. 2006. *Verb Classification, Case Marking and Grammatical Relationships in Amis*. PhD dissertation. Buffalo: State University of New York at Buffalo.

Pronouns in Discourse and Word Order Change in Cebuano

Michael Tanangkingsing

Department of English
National Taipei University of Technology

Keywords**:** Cebuano, discourse, pronouns, third-person forms, word order

In this study, I will observe pronominal occurrences, especially third-person forms, in narratives and conversations in Philippine languages and come up with patterns of uses of pronominal forms in these languages. Cross-linguistically, zero anaphora and bound pronouns are usually located at the top of the topicality scale (Givon, 1983), but languages differ in whether or not recoverable arguments can be omitted (Goldberg, 2004). For example, Mandarin Chinese generally allows the omission of pronouns. In contrast, English, as well as Tagalog (Nagaya, 2006), generally requires overt arguments. Similarly, in Cebuano, pronouns are employed mainly to refer to topical human participants, while zero anaphora is preferred for the less topical inanimate referents, as illustrated in (1). Moreover, pronominalization and zero anaphora occur in all argument slots, with A and S arguments tending toward pronominalization and P arguments slightly preferring zero anaphora. This is also obviously related to the fact that animate referents tend to occur in A and S slots, while inanimate entities in the P slot (where they are topical).

One of the two main observations made in this study is that based on spoken discourse data, the occurrence of two third-person pronominal forms in the same clause is dispreferred. In instances where two third-person referents occur in a transitive clause, there is a tendency to pronominalize the A referent, while the P referent will be anaphoric, as illustrated in (2) and Table 1. The other important finding is that due to the capability of pronominal arguments to be preposed to the front of the verb despite the fact that Cebuano is a verb-initial language, three patterns are observed, which lead us to suspect that Cebuano might represent a very early stage in the development from a Philippine-type language to an Indonesian-type language. First, specific verb-types (especially verbs of saying) are more likely to have pronominal A referents preposed. Second, transitive clauses with definite P referents are more likely to have pronominal arguments preposed. Third, pre-verbal elements in the verb complex allow pronominal forms to be preposed.

In conclusion, pronouns exhibit different types of behavior in discourse that traditional accounts do not show. The variations in pronominal forms and use contribute to pronouns playing a part in language change.

Data

(1) Cebuano: pronominal reference for animate referents

L *ma?ayo=**ka*** *kay* *na?a=na=man=**ka**=y* *anak*
good=2s.nom because exist=already=par=2s.nom=neut child
T *o=* *ma-* *ang* *ako-ng* *eldest* *23 years old*
bc fs ang 1s.nom-lk eldest 23.years.old
L *23=na*
23=already
T *ma-* *mag-* *n-anganak=na=**ko*** *ug* *beinte tres*
fs fs av-have.baby=already=1s.nom obl twenty.three
L *bata?=pa=man=**ka***
young=still=par=2s.nom
T *beinte dos=**siya*** *karon,* *kasi* *45=**ako*** *karon* *e*
twenty.two=3s.nomnow because 45=1s.nom now dm
tapos *kato* *ako-ng* *ika-duha*
then that 1s.poss-lk num-two
a= *three years-* *three years* *ang* *pagitan=nila*
fil three.years three.years ang space=3p.gen
kay *puro=man=**ko*** *caesarian*
because all=par=1s.nom caesarian
L: '(It's) good you already have kids.'
T: 'Yes, my eldest is 23 years old.'
L: '23 already?'
T: 'I gave birth (when I was) 23.'
L: 'You (look) still young.'
T: 'She's 23 now, because I'm already 45. Then, my second child, a= they're spaced three years apart, since I gave birth by caesarian.'

(4) Cebuano: excerpt from Pear Story

[1]*na-mu?pu?=siya* *ug* [2]*iya-ha-ng* *gi-butang*
intr-pick=3s.nom conj 3s.poss-def-lk pfv-place
sa *iya-ha-ng* *dako-ng* *sudl-an-an*
loc 3s.poss-def-lk big-lk inside-lv-nmz
[3]*tapos* *ni-na?og=siya,* [4]*iya-ha-ng* *gi-butang*
then av-go.down=3s.nom 3s.poss-def-lk pfv-place
sa *mora* *ug* *lamesa,* *pagka-human,*
loc like comp tablenmz-after
[5]*gi-trapu-han=usa?=niya* *ug* *tubaw*
pfv-wipe-lv=first=3s.gen ext handkerchief
[6]*tapos* *iya-ha-ng* *gi-butang* *sa* *basket*
then 3s.poss-def-lk pfv-place loc basket
ang *iya-ha-ng* *gipang-pu?pu?*
ang 3s.poss-def-lk pv-pick
'He picked (pear fruits) and he placed (them) (into) his big basket. Then, he came

down (from the tree), (and) he placed (them) on (something) like a table. Afterwards, he wiped (them) with a handkerchief. Then, he placed the (fruits that) he picked in the basket.'

Table 1. Referential tracking in excerpt (4)

clause	man	fruit	basket	table	handkerchief
1	S (pron)	E (zero)			
2	A (pron)	P (zero)	Loc (lex)		
3	S (pron)				
4	A (pron)	P (zero)		Loc (lex)	
5	A (pron)	P (zero)			Instr (lex)
6	A (pron)	P (lex)	Loc (lex)		

References

Givón, T. 1983. Topic continuity in discourse: an introduction. In Givón, T. (Ed.), *Topic Continuity in Discourse: A Quantitative Cross-language Study*. Amsterdam: John Benjamins, 1-42.

Goldberg, A. E. 2004. Pragmatics and argument structure. In Horn, L. and Ward, G. (Eds.), *Handbook of Pragmatics*. Oxford: Blackwell Publishing, 427-441.

Nagaya, N. 2006. Preferred referential expressions in Tagalog. *Tokyo University Linguistic Papers* 25: 83-106.

Do Corpora Work for All Levels? DDL with High and Low Proficiency learners

Asama Tasanameelarp[1] and Chonlada Laohawiriyanon[2]

[1] *Department of Languages Communication and Business*
Prince of Songkla University, Thailand
[2] *Department of Languages and Linguistics*
Prince of Songkla University, Thailand

Keywords**:** DDL, concordances, corpora, self-correction, grammar

Introduction

Due to the numerous benefits of computer corpora and concordances, there has been a growing and widespread use of these promising tools in computer-assisted language learning. One of the reasons for their increasing popularity is that language learners themselves can play an important role in learning inductively from authentic language information delivered to them in the form of concordance lines. Johns (1991a) termed this process "Data-Driven Learning" (DDL). However, some scholars have suggested that advanced learners benefit more from DDL than do learners at a lower level of language proficiency. Moreover, the majority of published studies in the use of corpora and concordances focus either on applications or on the effect of using corpora and concordances. Few studies, if any, have aimed to investigate the processes and strategies that learners adopt when dealing with these learning tools (Sun, 2003). Therefore, the study described in this paper investigated the use of corpora and concordances by two groups of Thai secondary school EFL students with low and high proficiency and their ability to self-correct grammatical errors and retain the knowledge gained by using corpora and concordances was examined. More specifically, this study explored the processes and the patterns of strategies used by Thai EFL learners, as well as their attitudes towards the materials they encountered.

The study adopted a quasi-experimental research design which was carried out with two groups of Thai grade 11 EFL learners over a period of 18 weeks. The instruments used included three error-correction tasks, a post-test, a retention test, and stimulated recall interviews. The teacher also took observation notes and conducted interviews. Prior to the experiment, the learners were asked to compose a story, prompted by a series of pictures. Then, the five most common types of errors at a word level were selected for the learners to correct and were used to design error-correction tasks and the retention test. During the experiment, the learners were trained to operate a concordancer, deal with concordances, induce patterns, and self-correct their writing.

The main findings established by paired sample T-tests and qualitative analysis can be summarized as follows:

1. There were significant differences in the ability of the two groups to self correct, with the students in the low proficiency group performing better in self-correcting than the high proficiency group in the post test while in the retention test the students in the high proficiency

group were more successful in self-correction than the low proficiency group.

2. The results from both groups show that the grammar categories of nouns, articles, and subject-verb agreement were those most successfully corrected in the correction tasks, the post-test, and the retention test. Verbs and prepositions were the grammatical types that were corrected the least successfully in all the data collection instruments

3. The qualitative results from the stimulated recall interviews revealed that different strategies were used by the high and low proficiency groups.

The results of this study indicated that the main factor influencing the learners' ability to identify information from concordance outputs and to construct knowledge capable of being used for self-correction was their prior grammatical knowledge. Moreover, the difficulty in self-correction experienced by the learners was found to be exacerbated by the complexity of the concordance lines and L1 interference.

References

Boulton, A. 2008d. DDL: Reaching the parts other teaching can't reach? In A. Frankenberg-Garcia (Ed.), *Proceedings of the 8th Teaching and Language Corpora Conference.* Lisbon, Portugal: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 38-44.

Johns, T. 1991a. Should you be persuaded: Two examples of data-driven learning. In T. Johns and P. King (Eds.), *Classroom Concordancing. English Language Research Journal*, 4, 1-16.

Sun, Y.C. 2003. Learning process, strategies and web-based concordancers: A case-study. *British Journal of Educational Technology*, 34/5, 601-613.

The Lexical Profile of Songs Used in the English Language Classroom

Friederike A. G. Tegge

School of Linguistics and Applied Language Studies
Victoria University of Wellington, New Zealand

Keywords: songs, pedagogical corpora, vocabulary learning, word frequencies, formulaic sequences, language teaching

Introduction

Many language teachers use songs in the language classroom for vocabulary teaching, for example to introduce or practice lexis. The following quote from an international online teacher survey with 467 participants (Tegge, in preparation) captures a common experience:

> I myself when I started learning English had listened to English songs to learn new vocabulary. Not only that it is enjoyable to listen to the songs …, it is a great way to help students learn new vocabulary …

However, some language instructors oppose the use of songs for teaching purposes or at least advise caution. One reason, they argue, is that the lexical content of songs is inappropriate for in-class use. In the same survey by Tegge, teachers voiced the following concerns:

> Often the vocab of readily available songs [is] at an advanced and poetic level...

> …songs are too difficult a thing … Especially because there [is] rarely any everyday/business language used in songs.

The Present Study and Literature Review

The present study investigates the lexical profile of songs used in English language classrooms. It is part of a broader research project investigating the opportunities for vocabulary learning afforded by songs. The analysis of a pedagogical song corpus is intended to shed light on the question whether the vocabulary of English songs – particularly pop songs but also folk and children's songs and songs written for language teaching purposes – renders them suitable or unsuitable for vocabulary learning. A particular focus of the corpus study lies on the investigation of the lexical content of the songs and the vocabulary knowledge needed to reach adequate comprehension. One approach is to use Nation's (2006)

BNC word-frequency lists. Results are compared to the lexical demands of other text formats such as TV programmes (Webb and Rogers, 2009b).

One factor affecting the retention of vocabulary is the number of encounters with a word. Webb (2007), for example, found that ten encounters with a previously unknown word in context lead to increased word knowledge. Thus, besides establishing a lexical profile, this study also investigates the repetition of words within and their range across songs.

In addition, an increasing number of studies highlight the importance of formulaic sequences for fluent and native-like language production and comprehension. Simpson-Vlach and Ellis (2010: 478), for example, observed that "… recent research in corpus linguistics … has established that highly frequent recurrent sequences of words … are not only salient but also functionally significant."

In the present study it is hypothesised that the presentation of formulaic sequences in a song format might, due to particular characteristics, aid their retention by language learners. The corpus is, thus, also analysed regarding the quantity of formulaic sequences.

Various corpus studies have investigated the vocabulary size necessary for learners to reach adequate comprehension, often defined as 98% vocabulary coverage (Hu and Nation, 2000), of various text genres. Schmitt and Adolphs (2003) found that the most frequent 5,000 words in the (spoken) CANCODE and the spoken part of the BNC achieve about 96% coverage. Nation (2006) found that for spoken texts knowledge of the 6,000 to 7,000 most frequent word families from Nation's (2006) BNC word-frequency lists are necessary, whereas for written texts knowledge of 8,000 to 9,000 word families is desirable. Webb and Rodgers (2009a) showed that an adequate comprehension of movies requires knowledge of 5,000 to 10,000 word families, whereas for TV programs Webb and Rodgers (2009b) suggest a necessary knowledge of 5,000 to 9,000 word families.

Research on the lexical demands of songs in the context of foreign and second language instruction is scarce. However, data from an earlier corpus of songs compiled by Murphey (1990) showed that 50 songs from a hit list are characterised by simple and semantically vague vocabulary. The type-token-ratio is lower than that of a first-year textbook. In addition, Murphey found that in an average song each word is repeated three times.

The Corpus

The present study differs from Murphey's (1990) in several respects. One difference is the size: The current corpus consists of over 600 songs. Another difference is the emphasis on the pedagogical character of this comparatively small and specialized corpus: It consists of songs recommended by teachers in an online questionnaire (Tegge, in preparation), as well as of songs used in textbooks and recommended by teachers on popular ESL/EFL websites. A further difference is the wider range of song genres. Sub-corpora have been analysed separately in order to capture and account for possible differences.

Several challenges rendered the compilation and analysis of the corpus more difficult than expected. For example, the number of repetitions of verses, choruses or single lines is of particular importance for this study, but they are often not provided in the lyrics. Also, a principled approach had to be found to deal with interjections (*yeah*, *oh no*) and lexically meaningless syllables (*hm*, s*halala*).

Preliminary Results

Results of a preliminary analysis indicate that around 88% of running words belong to the 1,000 most frequent word families in the BNC lists (Nation, 2006). To reach 98% coverage, knowledge of around 15,000 word families is necessary. Despite a high percentage of high-frequency words, it seems that a large vocabulary size is required to reach adequate comprehension. Implications of and explanations for these results will be discussed in the paper.

References

Adolphs, S. and Schmitt, N. 2004. Lexical coverage of spoken discourse. *Applied Linguistics*, 24 (4), 425-438.

Hu, M. and Nation, I. S. P. 2000. Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13 (1), 403-430.

Murphey, T. 1990. *Songs and music in language learning: an analysis of pop song lyrics and the use of song and music in teaching English to speakers of other languages*. Bern: European University Studies, Ser. 11, Education, Vol. 422.

Nation, I. S. P. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63 (1), 59-82.

Simpson-Vlach, R. and Ellis, N. C. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31 (4), 487-512.

Webb, S. and Rodgers, M. P. H.2009a. The lexical coverage of movies. *Applied Linguistics*, 30 (3), 407-427.

Webb, S. & Rodgers, M. P. H. 2009b. The vocabulary demands of television programs. *Language Learning*, 59 (2), 335-366.

Webb, S. 2007. The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28 (1), 46-65.

Patterns of Use of Category Ambiguous Words in Corpora and Coursebooks

Elaine W. Vine

School of Linguistics and Applied Language Studies
Victoria University of Wellington, New Zealand

Keywords**:** category ambiguity, general English, learner corpora, ELT coursebooks

Introduction

In recent decades, there has been significant progress in the development and application of general English corpora (e.g. O'Keeffe et al., 2007) and learner corpora (e.g. Granger, 2009) in English language teaching. There has also been some study of English language use in coursebook corpora (e.g. Meunier and Gouverneur, 2009).

The Study

In my study, I draw on the above three types of corpora to focus on category ambiguity in high frequency words, that is, where one word-form has different grammatical uses, for example, 'like' as preposition (1) and 'like' as verb (2):

(1) *i'm very interested that words **like** gestapo are suddenly coming into this business (dgi070).*

(2) *thats one of the reasons that i **like** them yeah (dpc265).*

(Holmes et al., 1998)

Category ambiguous words are known to be difficult areas for automatic taggers to handle, so my work includes manual analyses. I report on analyses and comparisons of the forms and functions of native English speakers' and English language learners' use of the following category ambiguous words:*about, as, can, down, like, over , round, so, up, will*. I present frequency data on the uses of these words in 1) British National Corpus and Wellington Corpora of Spoken and Written New Zealand English, 2) International Corpus of Learner English and Louvain International Database of Spoken English Interlanguage, and 3) a coursebook corpus which comprises the full texts of all books in two series of coursebooks, *New Headway* (NH) and *Cutting Edge* (CE), which are widely used with adult learners of English, ranging from beginners through to advanced learners.

Findings

There is variation in the extent to which the frequencies of occurrence in general and learner English corpora are consistent with each other and with pedagogical applications in coursebooks for different words. For example, Table 1 below shows the frequencies per 10,000 words of *up* and *down* used as adverb and preposition in the various corpora. (The frequencies of other parts of

speech for these two words, e.g. verb, noun, adjective, are close to zero.)

Up as adverb is used more frequently than *down* as adverb in all the corpora, but when *up* and *down* are used as preposition, the frequencies are comparable across the corpora. Looking in more detail at the adverb frequencies for *up* and *down*, spoken is higher than written for the BNC, but the reverse holds for the learner corpora. Furthermore, the learner corpora and the coursebooks show less frequent use of *up* and *down* as adverb than the BNC.

Table 1: Comparison across corpora of frequencies per 10,000 words of *up* and *down* used as adverb and preposition

up	**Adverb**	**Preposition**
BNC spoken	30	1.6
BNC written	17	0.7
LINDSEI	7	1.1
ICLEv2	12	0.2
NH series	16	1.6
CE series	14	1.5

down	**Adverb**	**Preposition**
BNC spoken	15	1.8
BNC written	8	0.9
LINDSEI	1.8	0.8
ICLEv2	2.7	0.2
NH series	4.7	1.0
CE series	4.3	1.1

There is also variation in the patterns of use of the other eight words in the data set. These sorts ofvariation raise issues about common pedagogical practices, which will be discussed briefly in the presentation.

References

Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer, K. (Ed.) *Corpora and Language Teaching,* Amsterdam: John Benjamins, 13-33.

Holmes, J., Vine, B. and Johnson, G. 1998. *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Meunier, F. and Gouverneur, C. 2009. New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In Aijmer, K. (Ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins, 179-201.

O'Keeffe, A., McCarthy, M. and Carter, R. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

A Contrastive Study of Corresponding Lexical Items in English and Chinese

Naixing Wei [1] *and Junying Zhou*[2]

[1] *Beijing University of Aeronautics and Astronautics*
[2] *Beijing University of Technology*

Keywords: equivalence, translation equivalents, semantic prosody, semantic preference, patterns

Semantic prosody — the use of language to express the speaker's attitude — has recently shed new light on contrastive lexical studies in a way that equivalence between bilingual lexical items can be established by examining both semantic and functional properties of words. While numerous previous studies have explored semantic prosody most extensively, few have attempted to investigate it from a bilingual perspective (notable exceptions include Tognini-Bonelli, 2002; Xiao and McEnery, 2006, etc.), or within a contrastive linguistic context, still less addressing how a node word's semantic preference and semantic prosody interactively contribute to an extended lexical unit on the basis of bilingual corpora evidence. The present study, drawing on insights from the Model of Extended Units of Meaning (Sinclair, 1996), sets out to examine equivalence between corresponding lexical items in English and Chinese by looking at their similarities in semantic preference and semantic prosody. We start with translation equivalents as shown in a bi-directional parallel corpus and further observe the preferential profile and prosodic norm of each translation equivalent in two comparable corpora, namely, the Modern Chinese Corpus (henceforth MCC) and the British National Corpus (henceforth BNC). A total of 25 pairs of translation equivalents have been investigated, with their respective mutual correspondence (Altenberg, 1999) measured and large numbers of concordances observed. The study indicate that similarities in semantic preference are insufficient for establishing lexical equivalents between Chinese and English, whilst it is the semantic prosodic similarity that plays a central role in establishing a bilingual lexical equivalent. With all the evidence examined, it seems that equivalence can hardly be found between single words of the two languages. But rather it is the various patterns in which words under comparison frequently occur that achieve avrying degrees of equivalence. Finally, the study addresses in more detail relationships between prosodic strength and equivalence. In the case of convergent prosodic norms, the less the prosodic strengths differ, the higher the degree of equivalence for a pair of translation equivalents. In the case of divergent prosodic norms, the less the prosodic strengths differ, the lower the degree of equivalence for a pair of translation equivalents. The contrastive study, among other things, calls for a need to shift undue focus on individual lexemes in traditional contrastive linguistics to larger units of meaning in the context of corpus research. It points to the

necessity of giving more weight to the semantic prosodies, as well as the patterns of words, in contrastive linguistic studies, translation studies and bilingual lexicography.

References

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hasselgård, H. and Oksefjell, S. (Eds). *Out of Corpora. Studies in Honour of Stig Johansson*. Amsterdam and Atlanta: Rodopi, 249-268.

Sinclair, J. McH. 1996. The Search for Units of Meaning. Textus Ix. 75-106.

Tognini-Bonelli, E. 2002. Functionally complete units of meaning across English and Italian. In Bengt Altenberg and Sylviane Granger (Eds). Lexis in Contrast. Amsterdam/ Philadelphia: John Benjamins, 74-95.

Xiao, R. and McEnery, T. 2006. Collocation, semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. Applied Linguistics. 27(1), 103–129.

What Can MAKE Tell us About Second Language Learning? – A Corpus-based Study of the Verb Complementation Patterns for MAKE

Qi Xu[1], Ming Yin[2] and Winfred Wing Fung Mak[3]

[1]*Department of English, The Chinese University of Hong Kong, Hong Kong*
[2]*English Department, Tianjin University, Tianjin, China*
[3]*Department of Education, The University of York, York, UK*

Keywords**:** verb complementation patterns, second language learning, corpus-based study, usage-based models

Introduction

For the last decade or so, there have been clear signs of linguistic researchers preferring to combine second language acquisition (SLA) and corpus linguistics, with their increasing awareness of the value of learner corpora. By comparing native speaker and learner corpora, the present paper investigates the verb complementation patterns for the high-frequency verb *make* used by native English speakers and Chinese EFL learners. It aims to find out the similarities and differences of the use of *make* by the two groups, and provide pedagogical implications for second language learning and teaching.

The examination of the verb *make* is actually not new in corpus-based studies. Similar research was also conducted by Altenberg and Granger , Liu and Shawand Altenberg , etc. What distinguishes this paper from the previous studies is mainly twofold: 1) the learner corpus in this study has been sub-divided into two different proficiency levels, i.e. beginning and advanced level, thus enabling the researcher to test the effect of language proficiency on SLA; 2) instead of examining the role of transfer as before, the author attempts to use the usage-based models to explain SLA phenomenon revealed from the corpus-based study.

To achieve the above objectives, two corpora have been adopted: the written sub-corpus of ICE-GB (*The British Component of the International Corpus of English*) and CLEC (*The Chinese Learner English Corpus*).

Results

Throughout the search of ICE-GB (*written texts*) and CLEC (*ST2*: beginning level, and *ST5&6*: advanced level), five verb complementation patterns for the verb *make* have been found, which are shown in Table 1. (Phrasal verbs like *make up*, *make out*, etc. are excluded.)

Table 1. Summary of the five verb complementation patterns for MAKE

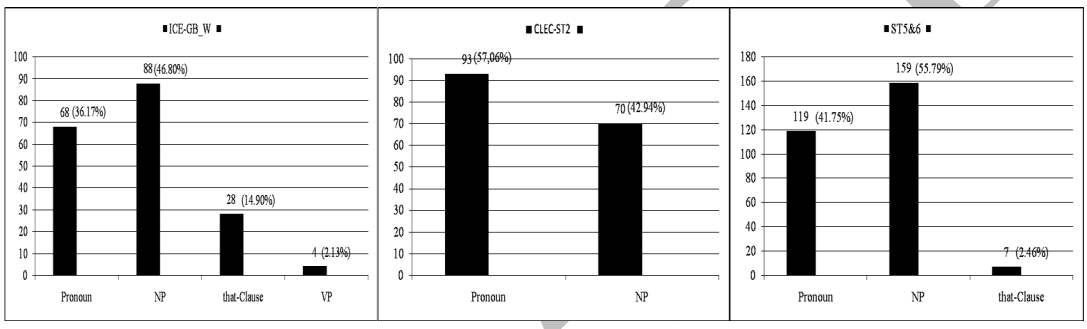
Patterns	***Basic structures***	***Examples***
1) Monontransitive	*make* + OD (direct object)	*I made a cake.*
2) Complex transitive	***make* + OD + OC** (object complement)	*They made me unhappy.*

3) Transitive	*make* + OD + nonfinite clause	*I made her laugh.*
4) Copular	*make* + SC (subject complement)	*You'll make a good leader.*
5) Ditransitive	*make* + IO (indirect object) + OD	*I made him a cake.*

It is found that Chinese EFL learners have a strong tendency to overuse the causative *make*, namely the complex transitive and transitive patterns. Further analysis is made concentrating on the use of **the complex transitive *make***, for the reason that most striking differences are detected between native and non-native speakers: 510 in ICE-GB, 807 in CLEC-ST2, and 674 in CLEC-ST5&6 (tokens/million words).

In the complex transitive pattern, two variables are examined – OD and OC. For the variable of OD, two major findings are discovered (see Figures 1-3). First, native speakers have a wider range of realizations as the OD, i.e., pronouns, noun phrases (NPs), *that*-clauses and verb phrases (VPs), while Chinese EFL learners' use only revolves around pronouns and NPs. Second, beginning learners use pronouns more frequently as the OD (57.06%), compared with the other two groups (native speakers: 36.17%; advanced learners: 41.75%).

Figures 1-3. Tokens of the OD in complex transitive *make* across three sub-corpora



In terms of the object complement, Learners overwhelmingly use Adjective phrases (AdjP) as the OC (beginning learners: 92.26%; advanced learners: 90.57%), but for native speakers, the AdjP only takes up 76.15%, and they also use a certain number of NPs and PPs to realize the form of OC, as shown in Table 2.

Table 2. Analysis of OC in three sub-corpora

Variable		ICE-GB_W			CLEC-ST2			CLEC-ST5&6	
OC	**AdjP**	166	***76.15%***	**AdjP**	155	***92.26%***	**AdjP**	269	***90.57%***
	NP	40	***18.35%***	**NP**	8	4.76%	**NP**	21	7.07%
	PP	11	5.04%	**PP**	4	2.38%	**PP**	7	2.36%
	Proform	1	0.46%	**AdvP**	1	0.60%			
	Total	218	100.00%	**Total**	168	100.00%	**Total**	297	9.43%

Brief Discussion and Conclusions

The major findings of the paper first suggest that beginning learners use a larger percentage of pronouns as the OD in comparison with native speakers and advanced learners. This, to some extent, proves the assumption of "pronoun islands" brought up by Jones et al. (2000) and reinforced by Childers and Tomasello , Ibbotson et al. (2010) and many others. Similar to children learning their mother tongue, pronouns also play an important role in forming linguistic schemas for L2 learners especially at the initial stages.

In addition, the results also show differences in the realizations of OC between native speakers and Chinese EFL learners. Although *make* is a high-frequency verb in English, even advanced learners are restricted to certain usages of *make*, for instance, structures like

"*make*+OC+AdjP". The reasons may be partly due to inadequate teaching and their insufficient input. Therefore, it is suggested that in SLA teaching, teachers and textbook compliers need to enrich the teaching materials not only by increasing the vocabulary store, but also by "fleshing out the incomplete or 'skeleton' entries" .

Acknowledgements

We owe special thanks to Prof. Gerald Nelson, who patiently guided us in extracting data from ICE-GB. Without his consistent support and insightful advice, we would not have accomplished this paper.

References

Altenberg, B. 2002. Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, and Petch-Tyson, S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Altenberg, B., and Granger, S. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics,* 22(2), 173-195.

Childers, J. B., & Tomasello, M. 2001. The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology,* 37(6), 739-748.

Ibbotson, P., Theakston, A., Lieven, E., and Tomasello, M. 2010. The role of pronoun frames in early comprehension of transitive constructions in English. *Language Learning and Development,* 7(1), 24-39.

Lennon, P. 1996. Getting "easy" verbs wrong at the advanced level. *International Review of Applied Linguistics in Language Teaching (IRAL),* 34(1), 23-36.

Liu, E. T. K., and Shaw, P. M.. 2001. Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics in Language Teaching (IRAL),* 39(3), 171-194.

Organizational Framework in English: The Case of "It is * that"

Suxiang Yang
School of Foreign Studies, South China Normal University,
Guangdong, China

School of Foreign Studies, Henan Polytechnic University,
Henan, China

Keywords**:** corpus linguistics, phraseology, organizational framework, variation

Introduction

Corpus linguistics has led to the development of a theory of language as phraseology (Sinclair, 1991; 2004; Cowie, 1998; Hunston and Francis, 1999; Granger and Meunier, 2008). There are three major categories of phraseology: meaning shift units, collocational frameworks and organizational frameworks (Warren, 2009). Organizational frameworks are the phraseologies which denote the ways in which organizational elements, usually the function words, may be co-selected by speakers and writers. For instance, 'I wonder * because', 'either * or', 'both * and' and 'whether * or', etc. This paper investigates the word variation and constituency variation of organizational framework "It is * that". The adjectives variation of the framework has been amply studied by traditional grammar and by corpus linguistics (Francis et al., 1996, 1998; Hunston and Francis, ibid.; Groom, 2005; Biber et al., 1999; Charles, 2000, 2004), but suffers from a lack of the other parts of speech and the constituency variation research.

Methods

The study reported here analyzed mainly qualitative meanings and functions of the structure. In this paper the fillers from one word to sequence of words of 'It is * that' were explored. The corpus used was BNC written texts and the corpus data were accessed using version 5.0 of Wordsmith Tools (Scott, 2009). The semantic classification of the fillers was based on Wmatrix (Rayson, 2009) and adjusted to this study. The research questions are: 1) when there is one filler in the slot, what types of words tend to occur in the frame? What are the parts of speech and what are the semantic categories of the types? 2) When the fillers are sequences of words, what are the word variation and constituency variation of the sequences? To answer these questions, I first investigated the parts of speech and the semantic classification of the one word fillers, and then explored the structure and meaning of the fillers from two words to five words sequences, and finally interpreted the motivation of the phenomenon.

Results

The results show that there are different patterns s and meaning of the fillers in the structure 'It is * that'. When there is one filler, many types of words could be used in the slot (*clear*, *likely*, *possible*). Further more, the parts of speech are adjectives (*clear, likely, possible*), past participles (*hoped, suggested, estimate*), present participles (*interesting, striking, encouraging*), and adverbs

(*simply, only, precisely*). The top three frequencies of the semantic categories are speech acts (*suggested, recommended, claimed*), importance (*important, vital, significant*) and thought (*assumed, believed, considered*). The motivation of the word variation is that link verb could be followed by adjectives, present participles and past participles according to traditional grammar and express speech act, evaluation and epistemic function. Adverbs can not be placed after the link verbs, however, so the pattern 'It is *adv*. that' is only to emphasize the adverb and 'It is *adv*. that' is equivalent to the adverb itself.

When the fillers vary from one word to five words they show different patterns.

1. If the fillers are two words sequences, the word sequence is an adjectives preceded by an adverb / negative word 'not' / 'a noun' / 'worth V-ing' / 'a phrase', etc. (*It is hardly surprising that*, *It is not surprising that*, *It is a pity that*, *It is worth noting that*, *It is no wonder that*).
2. If the fillers are three word sequences, it is often that the fillers are adverbs followed by 'to infinitive', or the passive 'to infinitive', or adverb followed by 'the case', etc. (*It is interesting to note that*, *It is to be hoped that*, *It is often/not the case that*).
3. if the fillers are four word sequences, the word sequences are the an adjective modified by an adverb before it, and then followed by 'to infinitive', or a phrase followed by an adjectives, or an adjectives followed by 'to infinitive consisting of two word', etc. (*It is also interesting to note that*, *It is by no means certain that*, *It is important to point out that*).
4. If the fillers are five word sequences, it shows a various patterns, but the comparatively more frequent word sequence is an adjective followed by 'to infinite with object' (*It is important to bear in mind that*, *It is difficult to avoid the conclusion that*, *It is difficult to escape the conclusion that*).

What's more, the motivation of the word variation and constituency variation of the different word sequences could be the fact that more words can express more detail and rich discourse of speech acts, importance and thought meanings and functions.

Conclusions

From the analysis, it shows that the organizational framework 'It is * that' reveals different words and constituency variations. For the word variation, different parts of speech and semantic categories of words could be filled in the slot. For the constituency variation, one to several words sequences could be put into the slot and revealed different patterns and function. The motivation is the requirement of the form and the meaning of the fillers and the organizational frameworks.

The significances of the study are: 1) theoretically, the study explored the discontinuous phrase, organizational framework, which is seldom researched in traditional phraseological studies, which usually focus on continuous phrase. 2) practically, especially for language teaching and learning, the results of the study could help the students to understand and produce the structure correctly.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Charles, M. 2000. The role of an introductory it pattern in constructing an appropriate academic persona. In P. Thompson (Ed.), *Patterns and perspectives: Insights into EAP writing practice*. CALS: The University of Reading, 45–59.

Charles, M. 2004. The construction of stance: *A corpus-based investigation of two contrasting disciplines*. Unpublished doctoral dissertation. UK: University of Birmingham.

Cowie, A.P. 1998. *Phraseology: theory, analysis and applications*. Oxford: clarendon Press.

Francis, G., Manning, E., and Hunston, S. 1996. *Collins COBUILD grammar patterns 1: Verbs*. London: HarperCollins.

Francis, G., Manning, E., and Hunston, S. 1998. *Collins COBUILD grammar patterns 2: Nouns and adjectives*. London: HarperCollins.

Granger, S. and F. Meunier. (Eds). 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam and Philadelpia: John Benjamins Publishing Company.

Groom, N. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4(3), 257-277.

Hunston, S. and G. Francis. 1999. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Rayson, P. 2009. *Wmatrix: a web-based corpus processing environment*. http://ucrel.lancs.ac.uk/wmatrix/

Scott, M. 2009. *Oxford WordSmith Tools Version 5.0*. http://www.lexically.net/downloads/version5/HTML/index.html

Sinclair, J. M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.

Warren, M. 2009. Why concgram?. In Greaves, C. 2009. *ConcGram 1.0: A phraseological search engine* – user manual: http://www.benjamins.com/jbp/series/CLS/1/manual.pdf (also on ConcGram CD).